Journal of Statistical Planning and Inference 155 (2014) 107–116

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Uniformly most powerful tests for simultaneously detecting a treatment effect in the overall population and at least one subpopulation

Michael Rosenblum *

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, 21205, USA

ARTICLE INFO

Article history:
Received 23 August 2013
Received in revised form 27 May 2014
Accepted 2 July 2014
Available online 10 July 2014

Keywords:
Familywise Type I error
Optimization
Personalized medicine

ABSTRACT

We take the perspective of a researcher planning a randomized trial of a new treatment, where it is suspected that certain subpopulations may benefit more than others. These subpopulations could be defined by a risk factor or biomarker measured at baseline. We focus on situations where the overall population is partitioned into two, predefined subpopulations. When the true average treatment effect for the overall population is positive, it logically follows that it must be positive for at least one subpopulation. Our goal is to construct multiple testing procedures that maximize power for simultaneously rejecting the overall population null hypothesis and at least one subpopulation null hypothesis. We show that uniformly most powerful tests exist for this problem, in the case where outcomes are normally distributed. We construct new multiple testing procedures that, to the best of our knowledge, are the first to have this property. These procedures have the advantage of not requiring any sacrifice for detecting a treatment effect in the overall population, compared to the uniformly most powerful test of the overall population null hypothesis.

1. Introduction

Planning a randomized trial of an experimental treatment can be challenging when it is suspected that certain populations may benefit more than others. For example, a metastudy by Kirsch et al. (2008) of certain antidepressants suggests that there may be a clinically meaningful benefit only for those with severe depression at baseline.

We focus on the case of two predefined subpopulations that partition the overall study population. For a given population, the mean treatment effect is defined as the difference between the mean outcome were everyone assigned to the treatment and the mean outcome were everyone assigned to the control. We develop procedures to simultaneously test the null hypotheses of no positive mean treatment effect for subpopulation one (H_{01}), for subpopulation two (H_{02}), and for the overall study population (H_{0*}). These hypotheses are defined formally in Section 3.3 below. For each of these null hypotheses, the alternative hypothesis is that there is a positive mean treatment effect for the corresponding population.

In some cases, there is a subpopulation for which a larger treatment benefit is conjectured. We call this the favored subpopulation, and refer to the other as the complementary subpopulation. Since it is usually not known with certainty before the trial that the treatment will benefit the favored subpopulation, planning a hypothesis test for it is important. Also, in trials where the overall population null hypothesis is rejected, it is of clinical importance to determine if the treatment benefits the complementary subpopulation, since there was more a priori uncertainty about the treatment effect for this

* Tel.: +1 410 614 9400.
E-mail address: mrosenbl@jhsph.edu.

http://dx.doi.org/10.1016/j.jspi.2014.07.001
0378-3758/

Table 1
Percent surviving 30 months, for the overall population and each subpopulation, by study arm. In each cell, the corresponding number of participants surviving 30 months divided by the number enrolled is given in parentheses.

	Overall population	Subpopulation one	Subpopulation two
Treatment arm	41% (96/235)	43% (62/143)	37% (34/92)
Control arm	32% (76/234)	35% (48/138)	29% (28/96)

group. Therefore, planning a hypothesis test for this subpopulation is also valuable. This motivates our interest in testing both subpopulation null hypotheses.

Our goal is to maximize power for simultaneously rejecting the overall population null hypothesis and at least one subpopulation null hypothesis. We give new multiple testing procedures that maximize this power, uniformly over all possible alternatives as defined in Section 3.6, for the case where outcomes are normally distributed. This optimality result may be of interest, since according to Romano et al. (2011), "there are very few results on optimality in the multiple testing literature". Our procedures require no sacrifice in detecting treatment effects for the overall population; that is, their probability of rejecting H_{0*} equals that of the uniformly most powerful test of H_{0*}, for any data generating distribution.

Our new procedures have a property called consonance. According to Bittman et al. (2009) "A testing method is consonant when the rejection of an intersection hypothesis implies the rejection of at least one of its component hypotheses". Consonance was introduced by Gabriel (1969), and subsequent work on consonant procedures includes that of (Hommel, 1986; Romano and Wolf, 2005; Bittman et al., 2009; Brannath and Bretz, 2010; Romano et al., 2011). Consonance is desirable since whenever an intersection of null hypotheses is false, it follows logically that at least one of the corresponding individual null hypotheses must be false as well. A non-consonant procedure may reject an intersection of null hypotheses without rejecting any of the corresponding individual null hypotheses. In our context, a non-consonant procedure would sometimes make claims that logically imply the treatment is superior to control in at least one of the two subpopulations, without indicating which one. To the best of our knowledge, our procedures are the first multiple testing procedures for our problem that are consonant.

Multiple testing procedures for the family of null hypotheses H_{0*}, H_{01}, H_{02} can be constructed using the methods of, e.g., Holm (1979), Bergmann and Hommel (1988), Maurer et al. (1995), Song and Chi (2007), Rosenbaum (2008), and Alosh and Huque (2009). However, none of these procedures is uniformly most powerful for simultaneously rejecting the overall population null hypothesis and at least one subpopulation null hypothesis, as defined in Section 3.6. We compare the power of our uniformly most powerful procedures to this prior work, in a simulation study in Section 6.

2. Example: randomized trial of treatment for metastatic breast cancer

Before giving formal details, we illustrate our hypothesis testing problem in the context of a randomized trial of trastuzumab for treating women with metastasized breast cancer (Slamon et al., 2001). As described in the abstract of Slamon et al. (2001), two types of patients were enrolled in the trial: those who had previously received an anthracycline as adjuvant therapy and those who had not. We call these types of patients subpopulation one and subpopulation two, respectively. This distinction between subpopulations is important since the concomitant treatments received by each during the trial were different. All subjects were randomly assigned to trastuzumab or placebo.

The results of Slamon et al. (2001) are based on survival data obtained 31 months after the last participant was enrolled; the range of follow-up was 30–51 months. We consider the outcome of survival at 30 months, and test the null hypotheses H_{0*}, H_{01}, H_{02}, that survival probability at 30 months is no greater under assignment to trastuzumab than under placebo, for the overall population, for subpopulation one, and for subpopulation two, respectively.

In Table 1, we give the percent surviving 30 months, for the overall population and for each subpopulation, by study arm, based on data in Fig. 2 of Slamon et al. (2001). The values of the z-statistics for the overall population, subpopulation one, and subpopulation two, are $Z_* = 1.89$, $Z_1 = 1.48$, and $Z_2 = 1.14$, respectively (rounded to two decimal places). The estimated correlation between Z_* and Z_1 is $\rho_1 = 0.79$, and the estimated correlation between Z_* and Z_2 is $\rho_2 = 0.62$. Formulas for these z-statistics and correlations are given in Section 3.5; in the above calculations, we substituted sample variances into these formulas.

We next introduce and demonstrate our new testing procedure. Let M^{UMP} denote the following multiple testing procedure for $\{H_{01}, H_{02}, H_{0*}\}$:

> Define S to be subpopulation 1 if $Z_1 - (3/4)\rho_1 \geq Z_2 - (3/4)\rho_2$, and subpopulation 2 otherwise. If $Z_* > \Phi^{-1}(0.95)$, reject H_{0*} and H_{0S}.

The procedure M^{UMP} strongly controls the familywise Type I error rate (defined in Section 3.4) at level 0.05. It is uniformly most powerful for simultaneously rejecting H_{0*} and at least one subpopulation null hypothesis, as defined in Section 3.6.

We apply M^{UMP} to the above data example. Since $Z_1 - (3/4)\rho_1 = 0.89$ and $Z_2 - (3/4)\rho_2 = 0.67$, the subpopulation S in M^{UMP} equals 1. Since $Z_* = 1.89$, which exceeds $\Phi^{-1}(0.95)$, our method M^{UMP} rejects H_{0*} and H_{01}. Existing multiple testing procedures based on the related work in the last paragraph of Section 1 (and which are defined in Section 6.1) either