Computational Statistics and Data Analysis 60 (2013) 111–122

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

An expected power approach for the assessment of composite endpoints and their components

G. Rauch *, M. Kieser

Institute of Medical Biometry and Informatics, University of Heidelberg, Im Neuenheimer Feld 305, 69120 Heidelberg, Germany

ARTICLE INFO

Article history:
Received 23 January 2012
Received in revised form 26 September 2012
Accepted 1 November 2012
Available online 17 November 2012

Keywords:
Clinical trial
Composite endpoint
Multiple testing
Non inferiority
Expected power

ABSTRACT

Composite endpoints are increasingly used in clinical trials, particularly in the field of cardiology. Thereby, the overall impact of the therapeutic intervention is captured by including several events of interest in a single variable. To demonstrate the significance of an overall clinical benefit, it is sufficient to assess the test problem for the composite. However, even if a statistically significant and clinically relevant superiority is shown for the composite endpoint, there is the need to evaluate the treatment effects for the components as, for example, a strong effect in one endpoint can mask an adverse effect in another. In most clinical applications, a descriptive analysis of the individual components is performed. However, the question remains what conclusion should be drawn from a trial where the composite shows a significant effect, but some component results which are not based on confirmatory evidence point in an adverse direction. Therefore, the first aim is to define an adequate multiple test problem of the composite and its most important components. Thereby, it might suffice to show superiority with respect to the composite and non-inferiority for the components to guarantee the clinical relevance of the result, as a slightly negative effect in one component might be acceptable as long as the total effect of all components is highly positive. To calculate the power for this multiple test problem, a number of strong assumptions on the effect sizes for the composite and its components as well as on the correlations between them are required. However, knowledge on these quantities is usually very limited and thus the choice of fixed parameter assumptions is based on a low level of evidence. The second aim therefore is to provide a more flexible power definition which takes the uncertainty about parameter assumptions into account. An expected power approach is proposed using prior distributions for the involved parameters. Thereby, the choice of the prior distribution reflects the level of evidence on the parameters. The expected power is evaluated for a range of scenarios and compared to the classical power for a fixed parameter setting. The new method is illustrated with a clinical trial example.

1. Introduction

Composite endpoints combine several events of interest within a single variable. The use of a composite can be motivated by different aspects. On the one hand, it increases the number of expected events which is meant to increase the power of the clinical trial. For example, in the field of cardiology trials, important endpoints like death or cardiac infarction are usually very rare events, so that the expected effect sizes are rather small and hence the required sample size is often too large to be realized. By combining several endpoints of interest into a composite, the effect size can be increased and thus the

* Corresponding author. Tel.: +49 6221 561932; fax: +49 6221 564195.
E-mail address: rauch@imbi.uni-heidelberg.de (G. Rauch).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.11.001

required sample size becomes smaller (Ferreira-Gonzáles et al., 2007). On the other hand, composite endpoints may also be applied when the clinical effect of interest cannot directly be captured by a single specific outcome measure but includes several endpoints (Lubsen and Kirwan, 1999). Instead of formulating a multiple test problem for several primary endpoints, these can be combined in a composite (ICH E9, 1999). For example, in the field of HIV treatment research, several surrogate parameters are often summarized within one variable to evaluate the efficacy of antiretroviral drugs (Wittkop et al., 2010). Thereby, the outcome of a composite endpoint can either be assessed as an 'event rate' or as 'time to first event', where the event is one out of several predefined variables of interest.

A major difficulty in the planning, analysis and interpretation of clinical trials with composite endpoints lies in the fact that the effect for the composite does not necessarily reflect the effects for the components (Freemantle et al., 2003; Freemantle and Calvert, 2007; Bethel et al., 2008; Kleist, 2010). Even if a statistically significant and clinically relevant effect in the composite has been observed, it may happen that the effects for some components are of very different magnitude or even point in the adverse direction. This is especially a problem if the single endpoints forming the composite are of different clinical relevance. It is therefore recommended in the literature not to combine endpoints of different severeness which, however, may be hard to fulfill in clinical application. As a consequence, in addition to the analysis of the composite endpoint an evaluation of the individual components is required (CPMP, 2002; Chi, 2005; Bethel et al., 2008). In clinical practice, the components are usually analyzed descriptively. However, in the case that a single component shows a negative effect in the descriptive analysis, conclusions on the efficacy or non efficacy of the new treatment are difficult, as in a descriptive analysis, the negative treatment effect might have occurred by chance. Consequently, it is desirable that the sample size for clinical trials that use composite endpoints as the primary variable should provide enough power not only to detect a clinically relevant effect for the composite but also to allow reliable confirmatory inference for the components forming the composite. The new approach is to define an adequate multiple test problem for the composite and its components. Thereby, it might seem desirable to show superiority for the composite and for all relevant components, however, this efficacy claim is a very strong one resulting in extremely high sample sizes. Depending on the specific situation at hand, it might suffice if all or some of the components show no major adverse effects, which is guaranteed by a non-inferiority claim, while the composite effect is tested for superiority.

Power calculation for such a multiple test problem is based on assumptions about the effect sizes for the composite and its components as well as their corresponding correlations (compare e.g. Song, 2009). As knowledge on these quantities is usually very limited, assumptions on fixed parameter choices are often not reliable and may differ considerably from the unknown true values. Therefore, a classical power calculation can be very misleading. A possible solution to this problem would be to calculate the classical power for a range of different plausible parameter settings and to base the sample size calculation on the worst case scenario. However, this approach might lead to unnecessarily high sample sizes. We propose to replace the fixed parameter assumptions by prior distributional assumptions, which better reflect the uncertainty about these parameters. By integrating the power function with respect to these prior distributions, an expected power can be defined (Brown et al., 1987; Spiegelhalter and Keith, 2004; Daimon, 2008). The weights are defined by the choice of the prior distributions which reflect the level of evidence on the parameters. The expected power is a more intuitive approach than the worst-case power approach as it incorporates the existing knowledge on the parameter assumptions. For example, if a large number of studies support specific parameter assumptions but a single study suggests more conservative assumptions, it would be inadequate to base the power calculation exclusively on these worst-case assumptions which correspond to a very small level of evidence. Instead, it seems reasonable to give a higher weight on the parameter assumptions from the majority of studies and only a small weight to the single study suggesting the more conservative parameter setting. Gillet (1994), Begum and King (2005) and Shao et al. (2008) discuss the advantages of an expected power approach over the classical frequentist power in different clinical trial situations.

The expected power can be interpreted as a semi-Bayesian power approach. Solutions for the Bayesian and semi-Bayesian sample size calculation have been widely discussed in the literature for the comparison of two proportions (Spiegelhalter and Freedman, 1986; Joseph et al., 1997; Katsis and Toman, 1999; Pham-Gia and Turkkan, 2003; Daimon, 2008) as well as for continuous outcomes (Cheng et al., 2010). However, so far no approach for a multiple binary test problem with correlated test statistics has been proposed.

In this work, we show how to define an adequate multiple test problem for the composite and its components in order to guarantee the clinical relevance of the result. To calculate the sample size for this multiple test problem, we present an expected power approach taking account of the uncertainty about the parameter assumptions. In Section 2, the hypotheses with the corresponding test statistics as well as the multiple test problem are introduced. The methods for frequentist and Bayesian power calculation are elaborated and the appropriate choice of the priors is discussed. Section 3 illustrates the proposed approach by a clinical trial example from the literature. The expected power is evaluated and compared to the frequentist power for a range of scenarios in Section 4. We conclude with a discussion.

2. Methods

2.1. Hypotheses

Consider a controlled clinical trial with a composite endpoint (CE) consisting of k components (EP_i, $i = 1, \ldots, k$). The sample size in the intervention group (I) is given by n, whereas the sample size in the control group (C) refers to m. Without