International Journal of Cardiology 175 (2014) 126–132

Contents lists available at ScienceDirect

International Journal of Cardiology

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

Easily applicable multiple testing procedures to improve the interpretation of clinical trials with composite endpoints

Svenja Schüler [a,*,1], Annegret Mucha [b,1], Patrick Doherty [c,1], Meinhard Kieser [a,1], Geraldine Rauch [a,1]

[a] Institute of Medical Biometry and Informatics, University of Heidelberg, Germany
[b] University of Bremen, Germany
[c] University of York, UK

ARTICLE INFO

Article history:
Received 20 December 2013
Received in revised form 25 April 2014
Accepted 30 April 2014
Available online 9 May 2014

Keywords:
Composite endpoints
Time-to-event
Multiple testing

ABSTRACT

Background: Cardiology trials often consider composite endpoints as primary efficacy outcomes thereby combining several time-to-event variables in a single time-to-first-event measure. The main motivation to use a composite endpoint is to increase the number of expected events thereby reducing the required sample size. However, interpretation may be difficult as the effect observed for the composite endpoint does not necessarily reflect the effects for the single components. To improve interpretation, it is therefore a current standard to analyze the individual components in a descriptive way. However, a descriptive analysis does not allow a statistical proof of concept. Therefore the gain in information is limited.
Methods: This paper systematically explores multiple testing procedures aimed at improving the interpretation of composite endpoints by confirmatory tests of the components. A simulation study demonstrates, on the basis of a real cardiology clinical trial example, the benefit of these easily applicable multiple testing procedures.
Results: By applying adequate multiple testing strategies to assess the components of a composite endpoint there is a high chance to get additional confirmatory evidence on the components without the need to increase sample size. With a moderate increase in sample size, a gain in evidence can often also be ensured with a predefined power.
Conclusion: The interpretation of composite endpoints can be improved by applying multiple testing procedures that assess the components. The methods discussed here are easy to apply and provide a substantial benefit for clinical interpretation of study results.

1. Introduction

In cardiology clinical trials, time-to-event endpoints often define the outcome variables of interest. Typical endpoints are given by death or specific causes of death, but also by nonfatal events like hospital admission, stroke or myocardial infarction. It is also a common practice to combine several events of interest into a single time-to-first-event variable, which is referred to as a composite endpoint [1,2]. By combining several event types into a composite endpoint, the number of expected events is increased and thus the required sample size is reduced [3]. Additionally, several outcome variables of interest can be simultaneously addressed without requiring adjustment of the type I error. The current standard is to restrict confirmatory analysis to the composite endpoint and to analyze the components only descriptively [4].

The problem of this approach is that the observed effect for the composite endpoint does not necessarily reflect the effects for the single components which can be different in magnitude. In the worst case, an adverse effect in one component is masked by a strong positive effect in another. Moreover, the single components often are of different clinical relevance [5]. In particular, death defines a component which clearly is more relevant for the patient than any other event of interest. Therefore, the interpretation of the composite endpoint becomes difficult if no information on the effects for the single components is taken into account. Descriptive analyses of the components are helpful but do not result in confirmatory evidence. Therefore, it is of high interest to develop methods which allow for a sound interpretation of the results by providing confirmatory information on the components also.

The present study aims to establish some easily applicable multiple test procedures which provide a gain in confirmatory information on the single components. The paper starts with a brief recall on the analysis of time-to-event data and on the interpretation of p-values. In order to give recommendations for medical experts on how to improve the interpretation of a specific study, different scenarios often met in clinical practice are addressed through the use of simulation studies which quantify the benefit of these methods in terms of power.

[*] Corresponding author at: Im Neuenheimer Feld 305, 69120 Heidelberg, Germany. Tel.: +49 6221 56 4177; fax: +49 6221 56 41 95.
E-mail address: schueler@imbi.uni-heidelberg.de (S. Schüler).

[1] This author takes responsibility for all aspects of the reliability and freedom from bias of the data presented and their discussed interpretation.

http://dx.doi.org/10.1016/j.ijcard.2014.04.267
0167-5273/© 2014 Elsevier Ireland Ltd. All rights reserved.

2. Methods

2.1. Analyzing time-to-event data

Time-to-event endpoints can be evaluated via standard survival analysis techniques such as the well-known Kaplan–Meier plot for graphical display, the logrank test as a statistical test for group comparison, or the Cox-model to incorporate covariates [6,7]. In a time-to-event setting, the instantaneous risk of experiencing an event at time t is the well-known hazard function. The standard logrank test compares the hazard functions between two groups by looking at the hazard ratio. The hypotheses for the logrank test are given by

$$H_0 : \lambda^C(t)/\lambda^I(t) \leq 1 \quad \text{versus} \quad H_1 : \lambda^C(t)/\lambda^I(t) > 1,$$

where C and I denote the group affiliation to the control and the intervention group, respectively. Thereby it is implicitly assumed that the hazard ratio is constant in time (proportional hazard assumption). For a composite time-to-first-event variable the hazard function is also referred as the all-cause hazard function, as it is based on more than one event type of interest. The all-cause hazard function is the sum of the hazard functions for the single components which are also called cause-specific hazard functions [8].

2.2. Interpretation of p-values and the idea of multiple testing

It is a current standard of reporting clinical trial results to provide the p-values from the logrank test for both the composite endpoint and for the individual components. In order to understand the benefit of a confirmatory analysis, it is necessary to recall the correct interpretation of p-values in such a context. Generally, a statistical test is regarded as significant if the corresponding p-value falls below a predefined boundary, called the significance level, which is usually given by 0.05. A p-value can either be significant or not, there exists no nearly significant or extremely significant result as the statistical test based on the p-value defines a binary decision rule. A statistical test is always constructed in the way that the type I error rate, that is the probability to falsely reject the null hypothesis although it is true, is no larger than the prespecified significance level α.

In cases where several hypotheses are simultaneously tested at significance level α the probability to falsely reject at least one of the null hypotheses is in general no longer controlled at this level. This phenomenon is called inflation of the type I error. When testing multiple hypotheses, it is therefore necessary to use a multiple testing strategy in order to control this false rejection probability by the so called ‘global significance level’. These testing strategies are referred to as multiple testing procedures.

Unfortunately, it is a widespread mistake and bad practice to consider a list of p-values and to mark all p-values smaller than 0.05 as ‘significant results’ as this approach ignores the multiple testing problem [9,10]. Indeed, the gain in information given by a list of p-values without incorporating the multiple testing aspect is minimal. If however, a multiple testing strategy has been specified in advance, then a list of p-values can give simultaneous confirmatory evidence on several test problems.

2.3. Composite endpoint scenarios commonly met in clinical practice

In order to analyze a composite endpoint and its components in a confirmatory way, the specific trial situation has to be considered. Although it is not possible to give optimal guidance for all clinical trial settings, there exist some typical situations often met in clinical practice. The remainder of the paper will give concrete recommendations for the following four scenarios:

1. The first scenario to be explored will be the situation of a composite endpoint consisting of individual components for which the expected effects are of different magnitude. As a typical example the DREAM trial showed a large effect in the component incidence of diabetes but only a small effect in the component death [11].
2. In a second scenario, a composite endpoint is considered where the components are either expected to be similarly affected by the new intervention or the component effects are difficult to predict. In many cardiologic trials different causes of death (e.g. sudden cardiac death, non-sudden cardiac death) are considered and combined into a composite. In this situation, it might be reasonable to assume that all causes of death are similarly affected.
3. The third scenario corresponds to the case where the composite endpoint contains one particularly harmful component, most often given by death, for which an effect in the wrong direction would not be acceptable, even if all other components show strong positive effects. The LIFE study used a composite endpoint of myocardial infarction, stroke and cardiovascular mortality [12]. Here, it would not be acceptable if cardiovascular mortality shows an adverse effect, even if myocardial infarction and stroke show strong positive effects.
4. As the last scenario, we consider the situation, where two different candidate endpoint combinations are under consideration in the planning stage and where it is not clear which of them is more meaningful and will show a higher effect. The two candidate endpoints can both correspond to composite endpoints or to single event endpoints. In this situation, a successful trial result can be based on the claim that at least one of the composite endpoints must be significant. The CAPRICORN Trial is a very illustrative example that shows how difficult it can be to determine an adequate primary endpoint in the planning stage. The original primary endpoint was all-cause mortality [13]. During a masked interim analysis, it was noted that overall mortality was lower than anticipated and the primary endpoint was changed to a composite endpoint of all-cause mortality or hospital admission for cardiovascular problems. The final study results, however, showed a significant effect for the mortality endpoint but no significance for the composite endpoint.

In the following, we will give recommendations on easy applicable multiple testing procedures for these four scenarios. The benefit of these methods will be illustrated in the section ‘Results’. A general overview on all scenarios and proposed solutions is provided below in Table 1.

2.4. Scenario 1: Hierarchical ordering for components with different effects

If a composite endpoint is used for which the effects of the components are expected to be of considerably different magnitude, an appropriate multiple testing strategy would be hierarchical testing. For illustration, consider a composite endpoint with three components. Without the loss of generality, component 1 is assumed to correspond to the largest effect followed by a smaller effect in component 2 and the lowest effect in component 3. Based on this assumption, the hierarchy of test hypotheses starts with the test of the composite endpoint, followed by the test for component 1 with the largest effect, subsequently the test for component 2 and finally component 3.

The left part of Fig. 1 shows the situation of hierarchically ordered components. If the individual hypotheses within a multiple test problem are hierarchically ordered, the test for an endpoint at an upper level defines a ‘gatekeeper’ for the subsequent tests. By this, only in cases where the composite null hypothesis is rejected, the null hypothesis for component 1 is tested. Otherwise any further null hypotheses in the hierarchy cannot be rejected. Subsequently only if component 1 is rejected the null hypothesis for component 2 is tested. Otherwise neither the null hypotheses for component 2 nor the null hypothesis for component 3 can be rejected. The procedure stops as soon as any individual null hypothesis is not rejected.

All tests are performed at the full predefined significance level α, which means that no adjustment to the local levels is necessary which proved a clear benefit [14]. However, it should be noted that the gain in information of this multiple testing strategy strongly depends on the ‘correct’ ordering of hypotheses by the magnitude of effects. If component 1 was expected to deliver a large effect in the planning stage but the observed effect in the study is small, the risk that the null hypothesis cannot be rejected is high. In this case none of the remaining components can be tested. Therefore, if the information on the expected effect sizes is limited in the planning stage, it might be better to use the multiple test procedure described in the next paragraph.

2.5. Scenario 2: Gatekeeping and Bonferroni–Holm for components with similar or unpredictable effects

In cases where the composite endpoint consists of components that are either similar in their expected effect sizes or the effect sizes are difficult to predict, it is not possible to arrange the components according to their effect sizes. In this situation, a suitable multiple testing procedure would be a combination of ‘gatekeeping’ and the Bonferroni–Holm procedure [14,15]. In the right part of Fig. 1 this multiple testing strategy is illustrated. Again it is necessary to reject the composite null hypothesis first before testing the components so that the composite endpoint defines a ‘gatekeeper’ for the component tests. However, once the composite null hypothesis is rejected, the components are no longer tested within a *predefined* order. Instead, the components are tested with the so called Bonferroni–Holm procedure [15].

Thereby the composite endpoint is tested at the predefined significance level α and once the composite null hypothesis has been rejected, the single components are tested in order of their observed p-values starting with the smallest. The smallest p-value is compared to the local adjusted significance level α/3 as there are three components to be tested here. If the null hypothesis corresponding to the smallest p-value can be rejected at this level, the null hypothesis with the second smallest p-value is tested at the local level α/2. If this hypothesis can also be rejected the remaining component can be tested at full level α. In the general case of k component null hypotheses to be tested, the adjusted local levels are given equivalently by α/k, $\alpha/k - 1$, …, α. Note that in contrast to the hierarchical ordering of hypotheses, an adjustment of the local levels is necessary in order to control the global type I error rate at level α. However, the advantage of this procedure is that no ordering has to be predefined in the planning stage.

2.6. Scenario 3: Intersection–union test in case of one particularly harmful component

So far, we have considered situations where the primary aim is to reject the null hypothesis for the composite endpoint but we try to address as many components as possible in a confirmatory manner. Now consider the case of a composite endpoint with one particularly harmful or patient-relevant endpoint. In such a situation, the primary objective is no longer to show a significant and relevant treatment benefit for the composite alone but to show *simultaneously* that at least no major adverse effect occurs for the main component. The primary aim is thus to show that the intervention group is superior to the control group in the composite endpoint *and* at least non-inferior with respect to the main component. Note that in most clinical trial situations due to the limited sample size, it will not be feasible to show superiority of the intervention with respect to both the composite endpoint *and* a single component. However, it might be sufficient to show that the new treatment beneficially affects the combination of all components and does not relevantly affect the main component in the opposite direction. To address the latter approach, a non-inferiority version of the standard logrank test is used to test the main component. To reflect this specific situation in an adequate multiple testing