Contemporary Clinical Trials 34 (2013) 257–261

Contents lists available at SciVerse ScienceDirect

Contemporary Clinical Trials

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

Short communication

Trial sequential analysis may be insufficient to draw firm conclusions regarding statistically significant treatment differences using observed intervention effects: A case study of meta-analyses of multiple myeloma trials

Branko Miladinovic [a,*], Ambuj Kumar [a], Iztok Hozo [b], Helen Mahony [a], Benjamin Djulbegovic [a,c]

[a] Center for Evidence Based Medicine and Health Outcomes Research, University of South Florida, Tampa, FL, USA
[b] Department of Mathematics, Indiana University Northwest, Gary, IN, USA
[c] H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA

ARTICLE INFO

Article history:
Received 9 July 2012
Received in revised form 18 December 2012
Accepted 20 December 2012
Available online 28 December 2012

Keywords:
Cumulative meta-analysis
Random error
Time-to-event outcomes
Bias
Trial sequential analysis
Spending functions

ABSTRACT

Trial sequential analysis (TSA) has been proposed as a method to assess the risk of random error in cumulative meta-analysis (MA), which increases due to repeated significance testing. The aim of TSA is to assist researchers from wrongly concluding treatment differences in the absence of a benefit (i.e. true versus false positive). Similar to monitoring boundaries applied in individual randomized controlled trials, recent literature has advocated the use of TSA for assessing the conclusiveness of results from MAs to determine the requirement for future studies in case of true positive results. While this may be desirable, we present empirical evidence from a recent systematic review to demonstrate that the use of TSA may lead to a premature declaration of statistically significant treatment difference, when further accumulated evidence suggested otherwise. Using all apparently conclusive MAs in multiple-myeloma, we empirically studied under what thresholds for the risk ratio reduction and power a true positive result becomes false positive. We recommend that the conclusion of significant treatment differences in cumulative MA should be weighed against acceptable thresholds regarding the type I error, power and *apriori* specified clinically meaningful treatment difference.

1. Introduction

Evidence from systematic reviews and meta-analysis (MA) of randomized controlled trials (RCTs) is considered most reliable for decision making related to therapeutic interventions. However, for the results from an MA to be relevant to the latest development in the field it requires updating as new trial results become available. Statistically, this implies that with the addition of a new trial repeated significance testing is performed, which increases the risk of random error and false positive results as the number of studies increases [1]. Similar to sequential group analysis in individual RCTs, sequential methods based on calculating monitoring bounds have been proposed to control for the risk of random error in cumulative MAs [2,3].

In the approach advocated by Pogue and Yusuf [3], for evidence obtained from MAs to be categorized as conclusive (true versus false positive or false negative), the number of participants, or optimal information size [4,5], should be at least as large as the sample size of a single optimally powered RCT. Wetterslev et al. [6] adjusted the method for bias and heterogeneity and labeled it trial sequential analysis (TSA). TSA provides the necessary sample size, monitoring and futility boundaries analogous to constructing interim monitoring boundaries for individual RCTs [7]. Recent studies have suggested that apparently conclusive evidence resulting from MAs may be inconclusive [6,8–17].

* Corresponding author at: Center for Evidence Based Medicine and Outcomes Research, Clinical and Translational Science Institute, Morsani College of Medicine, University of South Florida, 3515 E. Fletcher Avenue, MDT1200, Mail Code MDC27, Tampa, FL 33612, USA. Tel.: +1 813 396 9614; fax: +1 813 905 8909.
E-mail address: bmiladin@health.usf.edu (B. Miladinovic).

1551-7144/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.cct.2012.12.006

Some researchers may view the construction of monitoring boundaries for MAs as inappropriate, because meta-analysts do not usually have control over the generation of new evidence (i.e. enacting stopping rules) as they do in individual RCTs. However, meta-analysts do make recommendations based on pooled results [18], and as a result sequential methods can be viewed as a valuable tool for informing the decision making process and establishing conclusiveness of evidence without directly controlling future research [19]. Recent literature has advocated the use of TSA as a guide to determine requirements for future trials [10]. However, systematic assessment on the usefulness of TSA in deciding on the need for future studies has not been done. Accordingly, we have applied TSA on a cohort of systematic reviews in the field of multiple myeloma to assess whether results from TSA can be used to inform the need for future studies. This is especially important in light of recent work suggesting that evidence of efficacy based on a series of smaller trials may lower the error rates compared with single well-powered trial [20]. We chose multiple myeloma as a disease cohort based on our previously published work [21–23] and ongoing interest in this cancer field.

2. Methods

2.1. Information size and trial sequential analysis

In a cumulative MA, studies are added one at a time in a specified order (e.g. according to date of publication or quality) and the results are summarized as each new study is added. The conclusiveness of evidence is assessed at each stage, which is best visualized following the distinction in Brok et al. [8] and illustrated in Fig. 1. For the results to be conclusive (i.e. true positive) at 5% level of significance, cumulative trial Z-scores must cross both the $Z = 1.96$ lines and the monitoring boundary (curve B). If the lines do not cross both before reaching the required optimal information size, the result is considered false positive (curve A). Additionally, if the cumulative Z-curve crosses the futility boundary (curve D) [24], we can be confident that the treatment non-difference is not due to the lack of power, which is the case for curve C.

Briefly, the total number of observed patients in the cumulative meta-analysis is defined as the accrued information size. If the assumption is that the optimal information size (i.e. sample size) needed is at least equal to the sample size required in an individual RCT, given the pre-specified type I error α and power $(1-\beta)$ then the required optimal *apriori* anticipated information size (APIS) based on a pre-specified intervention effect μ and variance ν is defined as

$$\mathrm{APIS} = \frac{4\nu}{\mu^2}\left(Z_{\alpha/2} + Z_{\beta}\right)^2.$$

For binary outcomes and the event rates in the control and experimental groups p_c and p_e, $\mu = p_c - p_e$ and $\nu = p^*(1-p^*)$, where $p^* = (p_c + p_e)/2$. The *apriori* relative risk reduction (RRR) is defined as $\mathrm{RRR} = 1 - p_e/p_c$.

For time-to-event outcomes based on a pre-specified intervention effect HR_0 ($\mathrm{HR}_0 = 1 - \mathrm{RRR}$) APIS is defined as

$$\mathrm{APIS} = \frac{\left(Z_{\alpha/2} + Z_{\beta}\right)^2 (\mathrm{HR}_0 + 1)^2}{(1-w)(1-S)(\mathrm{HR}_0 - 1)^2}.$$

S is the average survival rate between the treatment and control arms and w is the expected censoring rate (i.e. loss to follow-up), given the pre-specified type I error α and power $(1-\beta)$ [25]. We note that a decrease in the assumed RRR means a higher value for APIS and more conservative monitoring bounds, which decrease the likelihood of finding a true positive result.

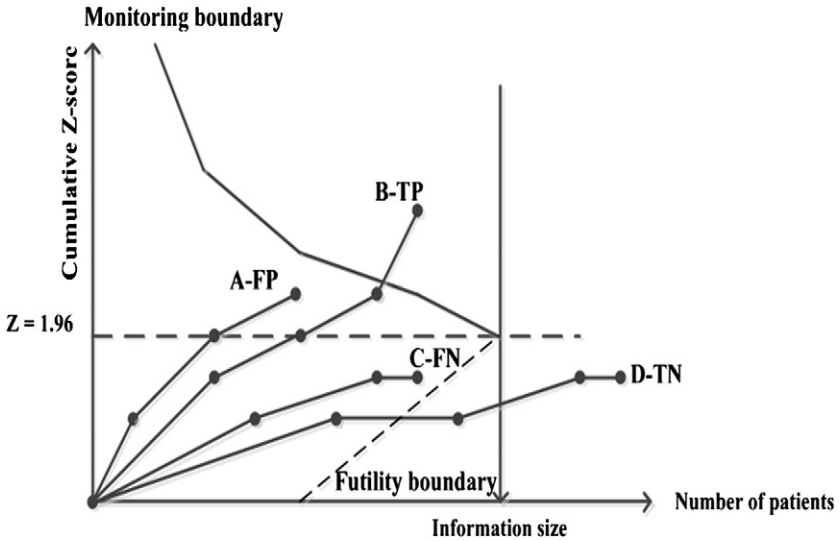

Fig. 1. Examples of the upper half of two-sided trial sequential analysis. For true significance, results must cross both Z and the monitoring boundary. Only B is a true positive. A is a false positive. C is a false negative because it included too few patients and is underpowered, and D is a true negative due to lack of a predetermined effect.