EUROMEDITERRANEAN BIOMEDICAL JOURNAL
2021,16 (25) 104-107
(*FORMERLY: CAPSULA EBURNEA*)

Commentary

TRIAL SEQUENTIAL ANALYSIS: THE EVALUATION OF THE ROBUSTNESS OF META-ANALYSES FINDINGS AND THE NEED FOR FURTHER RESEARCH

Filippo Sanfilippo [1], Luigi La Via [1], Stefano Tigano [1], Alberto Morgana [2], Valeria La Rosa [3], Marinella Astuto [1]

1. Department of Anesthesia and Intensive Care, A.O.U. Policlinico-San Marco, Catania, Italy
2. School of Anaesthesia and Intensive Care, Department of Medical and Surgical Sciences, "Magna Grecia", University, Catanzaro, Italy
3. Humanitas Istituto clinico catanese, Misterbianco (CT), Italy.

ARTICLE INFO

Article history:
Accepted 10 June 2021
Revised 29 June 2021
Published 15 July 2021

Keywords:
RCT, statistical error, systematic review, futility.

ABSTRACT

Randomized controlled trials (RCTs) may sometimes be underpowered to detect differences between interventions. Meta-analyses may reduce such risk by pooling together the available evidence from studies with similar inclusion criteria. However, this approach does not entirely rule out the risks of type-1 and type-2 statistical errors. Trial sequential analysis (TSA) is a statistical method developed for the assessment of the robustness of meta-analyses findings. Similar to the sample size calculation for a RCT, the calculation of the required "information size" for any TSA requires the outcome estimation (incidence or value) for the control and intervention groups, and the acceptable risks for type-1 and type-2 statistical errors (significance threshold and statistical power, respectively). Once performed, TSA graphically allows evaluation of whether the meta-analysis findings (cumulative effect size described in terms of Z-curve) are robust enough, or if there is need for further research. We briefly discuss the rationale and interpretation of TSA, identifying three main areas of the graph if the "information size" has not yet been reached. No further research is needed when the Z-curve crosses the adjusted significance thresholds or stands in the futility area. Conversely, if the Z-curve sits between adjusted significance thresholds and the futility boundary, more research is needed to establish differences between the interventions. In summary, similar to the interim analyses performed during any RCTs, the TSA graph can be helpful for scientists to understand if a meta-analysis finding is robust, and if further research would be desirable or futile on that topic.

1. Introduction

Multicenter RCTs are the gold standard of research to evaluate differences between interventions. However, for several reasons their results may be underpowered for the outcomes of interest, leaving uncertainty regarding the validity of the findings. Meta-analyses of RCTs represent the top of the pyramid for evidence-based medicine, with the advantage of increasing the sample size. Indeed, meta-analyses have been developed with the target of pooling together all the available evidence from studies with similar inclusion criteria. The consequently larger sample size reduces the risks of type-1 and type-2 statistical errors.

It has been estimated that over 10 systematic reviews are published daily in medical literature [1]; however, a large proportion of these meta-analyses are redundant, misleading and poorly conducted [2].

* *Corresponding author*: Filippo Sanfilippo, filipposanfi@yahoo.it
DOI: 10.3269/1970-5492.2021.16.25
All rights reserved. ISSN: 2279-7165 - Available on-line at www.embj.org

Producing positive or negative findings from a systematic review and meta-analysis does not necessarily mean these results are robust and valid. Indeed, it is possible that these findings still encounter type-1 and type-2 statistical errors. In brief, type-1 statistical error (α) happens when a true null hypothesis is rejected; in other words, a significant difference is reported between two interventions, but such a difference is false. Type-2 error (β) is encountered when no difference is reported, but a significant difference between two interventions exists (failure to reject a false null hypothesis).

An attractive statistical method has emerged to evaluate the robustness of the findings of a meta-analysis titled trial-sequential analysis (TSA). It combines conventional meta-analytical approach with the introduction of monitoring boundaries. Such boundaries represent thresholds for determining the significance of results; they are developed according to multiple testing and to the amount of information already available in the meta-analysis. Our brief commentary describes basic principles underlying the TSA and its interpretation.

2. A parallel between TSA and interim analyses in RCTs

A practical approach to the world of TSA could be the use of an analogy with the interim analyses conducted during RCTs. It is a good and standard practice to perform interim analyses at predefined time-points during the RCT, before the trial has reached its sample size. The sample size calculation for any RCT is performed with the aid of four pieces of information:

a-Event incidence in the control group;

b-Anticipated effect size of the intervention;

c-Acceptable risk of type-1 statistical error (usually<5%);

d-Requested statistical power (normally at least 80%).

Interim analyses are conducted for several reasons, most of which are ethical and financial. Indeed, if an intervention is clearly superior (or inferior) to the other one studied (or to a placebo), the RCT should be stopped as it would be unethical to randomize patients to the least effective intervention. On the other hand, it would be a waste of time and resources to continue the enrollment in a RCT where the two interventions being studied have equipoise effects; continuing the RCT will be futile, as there are no chances that collecting further data would produce a statistically significant effect.

A TSA may be considered a sort of interim analysis for the meta-analysis. Indeed, from the interpretation of the TSA, scientists may be able to understand if a positive or a negative finding is true (or there is risk for type-1 or type-2 statistical error), and if more research would be desirable or futile.

In fact, TSA aims at saving time and resources, minimizing participants' exposure to the inferior treatment, and avoiding unnecessary research when the findings are robust or when it is futile to further investigate differences between interventions. Finally, TSA approach could be useful to inform scientists on the required number of participants to "fill the gap".

3. The elements of a TSA

The same four pieces of information needed to calculate the sample size in RCTs (event rate, estimated effect size, significance level and statistical power) are necessary to perform a TSA.

Using the same assumptions to evaluate the required "information size", the TSA also creates "adjusted" significance thresholds in order to reduce the risk of misinterpreting random error, which increases when data are sparse and "information size" has not been reached.

Five elements characterize a TSA and its relative result (Figure 1):

a. the cumulative Z-curve (blue color) representing the effect size, which may (or may not) favor one intervention over another. It is a summary test statistic of all the included studies, and a new Z-value is calculated each time a new study is added;

b. the required "information size" calculated as just described, and identified by the vertical red line on the right end of the figure;

c. the conventional significance boundaries, which depend on the decided α level (assuming $\alpha=0.05$; these two boundaries are set at ±1.96 standard deviations, and identified as horizontal lines in brown color);

d. the curved monitoring boundaries (red color, also referred to as TSA-adjusted significance thresholds) which, as compared to the conventional significance boundaries, apply stricter thresholds when the accumulated "information size" is lower than the required one. Basically, for small accumulated "information size", monitoring boundaries will be large, meaning that a greater difference between intervention is needed when the number of observations is low;

e. the futility boundaries (red color lines) that identify a triangular area located in the right-sided half of the TSA figure. When the Z-curve sits in this area further research would be futile, as the increase in the accumulated "information size" has virtually no chances to bring the Z-curve outside the significance thresholds.

As shown in Figure 1, the presence of the above-mentioned lines separates several areas in the TSA graph. Assuming the "information size" has not yet been reached, the Z-curve may sit in three different areas. The first possibility is that the difference between interventions is large enough and the Z-curve crosses the adjusted significance thresholds; in this case the difference is statistically significant and robust enough so that no more research is needed (light blue area). When the Z-curve stands in the futility area (light yellow color, as in the case of Figure 1), there is no difference between interventions, and it is very unlikely that more research would change this result. The third area (light green) is the zone that identifies the need for further research as the Z-curve sits between the futility area and the adjusted significance thresholds; in such cases, regardless of statistical significance according to the conventional boundaries, the meta-analysis results are not robust and more research is needed to ascertain the differences between the two interventions.

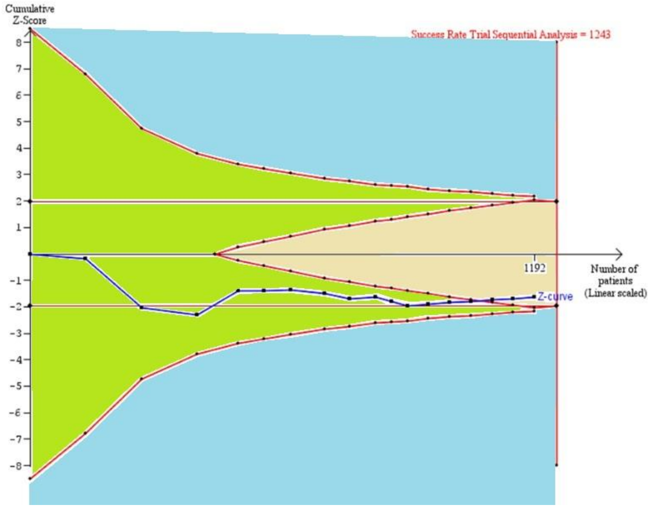

Figure 1. Example of Trial Sequential Analysis evaluating the time to tracheal intubation comparing individuals wearing standard uniforms as compared to personal protective equipment (author's own data). In the case shown in figure, there is no differences between interventions; moreover, the Z-curve crosses the "futility boundaries" (standing in the light-yellow area) and therefore, no more research is needed even if the required "information size" have not been reached. The light blue area is located outside the "adjusted significance thresholds"; when the Z-curve stands in this area, the difference between interventions is significant and the result is robust. When the Z-curve sits in the green area, more research is needed to establish if a difference exists between two interventions.

4. Heterogeneity and diversity

For multi-center RCTs, it is usual practice to correct for between-centers variations. In the case of meta-analyses, an estimation of the heterogeneity among the included studies (I2) is usually performed to account for differences in populations, designs, interventions, and effect size. Similarly, the TSA assesses the heterogeneity (or diversity, D2) as the measure to account for between-trial variations, similar to the I2 used in a conventional meta-analysis [3]. The greater the heterogeneity, the lower the precision of the results.

5. Sparse data, repeated testing and spurious findings

Two features identify meta-analyses as more predisposed to incorrect findings: presence of sparse data (i.e. low numbers of included trials or small number of events)[4,5], or frequent updates (repeated significance testing increases the likelihood of type-1 errors)[6]. It has been suggested that 10%-30% of interventions may be falsely reported as beneficial (or useless) by meta-analyses due to type-1 error[5,7].

6. Limitations of TSA

TSA can be applied to meta-analyses of both randomized and non-randomized studies; similarly, TSA may investigate dichotomous parameter as well as continuous data; however, continuous data can be assessed only via mean difference (standardized mean differences is not utilizable).

One issue when performing a TSA is the author's choice on the anticipated effect size of the intervention. Ideally, authors should specify preventively in their protocol the criteria that will be adopted for the TSA. A commonly adopted approach is to use the relative risk reduction or the mean difference according to the weighted averages gathered from the included studies. This approach has the value of balancing the effect size according to those reported by the included studies; conversely, TSA conducted with this approach may use differences that are not clinically meaningful. A different approach could be to use the differences reported by the largest included study, if it is considered representative. Another limitation is the typical author's decision to perform the TSA for their primary outcome only, meaning that the risk of falsely positive or negative findings persists for the secondary outcomes [8]. The TSA is not favored by all scientists, and the most recent Cochrane Scientific Committee Expert Panel recommended against routine use of TSA[9], arguing that sequential methods (such as the TSA) cannot accommodate multiple different thresholds for different outcomes.

7. Conclusions

Performing TSA is an increasingly adopted approach to address the robustness of meta-analyses findings, especially when sparse data are pooled. We provided a brief summary on the interpretation of TSA results, while bearing in mind some of its limitations. For those interested in learning more about the underlying principles of TSA, the manual can be freely downloaded at this website www.ctu.dk/tsa together with the software developed to perform the TSA.

References

1. Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? PLoS Medicine 2010; 7: e1000326.
2. Ioannidis JP. The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. Milbank Quarterly 2016; 94: 485–514.
3. Wetterslev J, Thorlund K, Brok J, Gluud C. Estimating required information size by quantifying diversity in random-effects model meta-analyses. BMC Medical Research Methodology 2009; 9: 86.
4. Brok J, Thorlund K, Gluud C, Wetterslev J. Trial sequential analysis reveals insufficient information size and potentially false positive results in many meta-analyses. Journal of Clinical Epidemiology 2008; 61: 763–9.

5. Borm GF, Donders AR. Updating meta-analyses leads to larger type I errors than publication bias. Journal of Clinical Epidemiology 2009; 62: 825–30.
6. Chandler JS, Clarke M, Higgins JPT. Cochrane methods. Cochrane Database of Systematic Reviews 2012; 1: 29–35.
7. Imberger G, Thorlund K, Gluud C, Wetterslev J. False-positive findings in Cochrane meta-analyses with and without application of trial sequential analysis: an empirical review. British Medical Journal Open 2016; 6: e011890.
8. Heesen M, Klimek M, Hoeks SE. Restrictive or responsive? Outcome classification and unplanned sub-group analyses in meta-analyses. Anaesthesia 2018; 73: 279–83.
9. Cochrane Scientific Committee. Should Cochrane apply error-adjustment methods when conducting repeated metaanalyses? 2018. https://methods.cochrane.org/sites/default/files/public/uploads/tsa_expert_panel_guidance_and_recommendation_final.pdf (accessed 25/03/2019).