Trial sequential analysis in meta-analyses: A clinically oriented approach with real-world example

Jeffrey Shi Kai Chan, MBChB,a,b and Amer Harky, MBChB, MRCS, MSc

Supplemental material is available online.

Ranked high on the hierarchy of evidence, meta-analyses are important summaries of the existing literature, often influencing clinical practice and guidelines. However, as evidence accumulates, random errors may at times lead to spuriously significant results, and an increasing frequency of statistical testing by meta-analyses increases the likelihood of such being reported; that is, type 1 error. Therefore, trial sequential analysis (TSA) has been developed to adjust for this increase in type 1 error. TSA is conceptually similar to sequential interim analyses of randomized controlled trials, where trial results are tested at regular intervals to determine whether a certain difference due to the intervention, or the lack thereof, has been conclusively demonstrated by existing data, allowing the trial to be terminated.

In TSA of meta-analyses, data from each study is analogous to data added during each sequential interim analysis. Statistical techniques are used to adjust and increase the threshold for statistical significance (z-score threshold; conventionally corresponds to a P value of .05) based on the effect to be observed, incidence of outcome in the control arm, information size (a statistical measure of the amount of data analyzed), and heterogeneity. Simply put, a large effect (eg, 90% relative risk reduction [RRR]) observed from data with minimal heterogeneity requires smaller information sizes for the observation to be conclusive than a small effect (eg, 10% RRR) observed from data with significant heterogeneity. As such, TSA may be used to determine whether certain effects (ie, differences between 2 arms, such as a 25% RRR) may be considered conclusive in a meta-analysis. Results found to be conclusive may not require much further investigation, whereas premature and spurious conclusions based on unadjusted significant results can also be avoided.

Meta-analyses with TSA and post hoc TSA have become more common in recent years, having been incorporated in some Cochrane reviews as well. It is therefore important for clinicians to understand the principles, interpretation, and limitations of TSA. In this article, we aim to illustrate these with a post hoc TSA using data from a meta-analysis by Woldendorp and colleagues. The meta-analysis shows an important association between subclinical valve thrombosis (SCVT) of transcatheter aortic valve implantation and stroke, and the superiority of oral anticoagulants (OAC) to single (SAPT) or dual (DAPT) antiplatelet therapy in preventing SCVT.

To perform the TSA, study-level summary data of all meta-analyzed comparisons were obtained from the published Forrest plots, with odds ratios as summary estimates in accordance with the choice by Woldendorp and colleagues. Single-arm studies were excluded. Two-sided z-score thresholds were adjusted using the O’Brien-Fleming α-spending function with 2-sided 5% type 1 error.
and 80% power. The DerSimonian-Laird random effect model was used. Control arm incidences were calculated using event rates from all studies. Information sizes were estimated from all available statistical information (Fischer information). If that was impossible, sample sizes were used instead. In such cases, information size estimations were repeated using incidences of outcome event (event sizes) to ensure the robustness of calculations. RRR values were estimated from all studies. To demonstrate the interpretation and utility of TSA performed using prespecified RRR values, TSA was also repeated with RRR values deemed appropriate for the comparison made.

Studies reporting no event were handled by adding a constant (1) to both the event and sample counts. All statistical analyses were performed using TSA software version 0.9.5.10 Beta, Copenhagen trial unit. All TSA data files are available online (Online Data Supplement 1-3).

TSA generates graphical outputs (Figures 1-4). The y-axis represents the cumulative z-score of included studies, which is a measure of the statistical strength of the cumulative evidence, whereas the x-axis represents information size. The blue line is the z-score curve, which is a plot of the cumulative z-score with study data added and plotted chronologically. These data points may


represent the sample size, event size, or a combination of both of these and the number of trials (also called Fischer information), depending on the specification of the analysis. Generally, Fischer information is preferred and superior to the other options. The brown, dotted lines represent the conventional statistical significance threshold for 2-sided -score values of ±1.96 (these are optional and not always present in TSA plots). The red lines represent the adjusted boundaries (Figure 1)—a pair of outer, oblique red lines form the statistical significance boundaries, whereas the inner oblique lines form the futility boundaries. Crossing either of the statistical significance boundaries imply that the existing, statistically significant data are conclusive; similarly, crossing either of the futility boundaries imply that the existing, statistically insignificant data are conclusive. The final vertical red line represents the required information size (RIS)—upon reaching this information size, all observations are said to be stable because effects are unlikely to fluctuate significantly even upon addition of new data, and a conclusion has therefore likely been reached. The statistical significance boundaries become narrower and approaches the conventional statistical significance threshold as information sizes increases, reflecting that the significance threshold is penalized and adjusted more heavily when the information size is smaller. Figure 1 shows the TSA results comparing stroke rates in patients with SCVT or normal leaflet motions (NLM) (stroke rate in the latter, 2.75%). Sample sizes were used...

Similar observations were made when the TSA was performed using event size instead, suggesting that the calculation was robust.

Figure 2 shows the TSA results comparing OAC to SAPT or DAPT for preventing SCVT (SAPT or DAPT arm incidence, 18.71%). Again, the \( z \)-score curve crossed the RIS, suggesting that the observed effects could be considered conclusive. With an observed RRR of 68.6%, the TSA was repeated with an RRR of 20%.

Similarly, the \( z \)-score curve crossed the RIS, indicating that the existing evidence is sufficient to conclude that there is at least a 20% reduction in SCVT incidence in

patients taking OAC, compared with those taking SAPT or DAPT.

Figure 3 shows the TSA results comparing SAPT to DAPT for preventing SCVT (DAPT arm incidence, 16.06%). Again, this was performed using sample sizes as Fischer information could not be used due to too little information (−63.12%). In Figure 3, the z-score curve did not cross any boundary. This indicates that the observed effect was not conclusive, and more studies are required before any conclusion could be made. With an observed RRR of −16.4%, TSA was repeated with an RRR of −20%, with similar results that indicate that the current evidence is insufficient to confirm or rule out a 20% increase in SCVT risk in patients taking SAPT compared with those taking DAPT. TSA could also be repeated with an RRR of −33.3%, as shown in Figure 4, which shows that the z-score curve crossed the futility boundary. This leads to the conclusion that the existing evidence is sufficient to show that SAPT use is not associated with an increase in SCVT risk of 33.3% or more compared with DAPT. Similar observations were made when the TSA was performed using event size instead.

Although the conclusions of TSA may be attractive, one must bear in mind the limitations of TSA. Most importantly, virtually all limitations of meta-analyses are applicable to TSA as well, because TSA makes use of data from meta-analyses. Any significant risk of bias, including publication bias, can affect TSA results significantly. Therefore, TSA is ideally performed on study data with minimal risk of bias. Nonetheless, most meta-analyses inevitably include studies with high risks of bias. To ameliorate this issue, it is possible to estimate effect sizes for TSA only from studies with low risks of bias. On a similar note, significant heterogeneity also affects the validity of TSA results—in the above TSA example, whereas the OAC was shown to be conclusively superior to SAPT or DAPT for preventing SCVT, the results should be interpreted with caution as significant heterogeneity was detected ($I^2 = 68\%$). More detailed analysis, including subgroup analyses or meta-regression, may be performed to delineate the underlying cause of such heterogeneity and refine the conclusion. An alternative would be to perform TSA on outcomes without significant heterogeneity, because the heterogeneity would need to be addressed before the evidence could be considered conclusive and definitive anyway.

As useful as TSA may be in preventing spurious conclusions, TSA have also been criticized for being too conservative. This relates to the RRR selected for analysis—choosing RRR values that are too conservative easily leads to results showing that existing evidence is insufficient to confirm or rule out the chosen level of RRR. Indeed, this is mostly arbitrary in most cases and relies heavily on the clinical expertise and experience of the author. As shown in the above example comparing the stroke rates in patients with and without SCVT, it would be almost meaningless to repeat the TSA with a RRR of −20%, which is a common prespecified RRR value used in TSA: the z-score curve does not cross any boundary when such a small RRR is used, and the resultant conclusion that the current evidence is insufficient to confirm or rule out a 20% increase in stroke risk with SCVT is unhelpful given the strong, conclusive evidence that there is at least a 100% increase in stroke risk. Such difference in TSA results may seem counterintuitive, but is simply a result of the previously mentioned principle that more data are required to conclusively demonstrate a small effect size compared with a large effect size. The chosen RRR therefore directly influences the results and conclusion of TSA, and clinical relevance and appropriateness of any chosen RRR must be considered when interpreting TSA.

CONCLUSIONS

TSA of the meta-analysis by Woldendorp and colleagues revealed that existing evidence is sufficient to conclude at least a 100% increase in risk of stroke in patients with SCVT compared with those with NLM, as well as that OAC use is associated with at least 20% reduction in SCVT risk compared with SAPT or DAPT use. However, existing evidence is insufficient to rule out or confirm a 20% increase in SCVT risk in patients who take SAPT, compared with those taking DAPT; this was despite the evidence being sufficient to rule out an increase in SCVT risk of 33.3% or more. Our analysis demonstrated the key utility of TSA, an evolving tool that may be expected to play increasingly important roles in meta-analyses. Clinicians should familiarize themselves with TSA, and note the limitations of TSA when interpreting its results. When used and interpreted appropriately, TSA can be a powerful tool capable of assessing the conclusiveness of meta-analytical findings.

Conflict of Interest Statement

The authors reported no conflicts of interest.

The Journal policy requires editors and reviewers to disclose conflicts of interest and to decline handling or reviewing manuscripts for which they may have a conflict of interest. The editors and reviewers of this article have no conflicts of interest.

References


