Trials

COMMENTARY Open Access

Lack of reproducibility of trial sequential analyses: a meta-epidemiological study

Xing Xing[1], Yining Wang[2], Yipeng Wang[3], Mohammad Hassan Murad[4] and Lifeng Lin[5*]

Abstract

Systematic reviews and meta-analyses are essential tools for synthesizing evidence from multiple studies. Recently, trial sequential analyses (TSAs) have gained popularity as a component of meta-analyses, helping researchers dynamically monitor evidence as new studies are incorporated. This article introduces a meta-epidemiological study aimed at evaluating the reproducibility of TSAs within systematic reviews published in 2023. Two independent investigators assessed and reproduced the main TSA for each included systematic review. Our search in PubMed yielded a convenience sample of 98 systematic reviews. Only 28% (27/98) of the included TSAs provided sufficient data to calculate the required information size, an essential element for assessing statistical power and conducting TSAs. Among these, 81% (22/27) provided the necessary data to determine decision boundaries and Z-curves in TSAs. Overall, full reproducibility was achieved for only 13% (13/98) of TSAs. Specifically, for binary outcomes, 65% (47/72) of TSAs failed to report event rates in control groups, and 44% (32/72) did not report relative risk reductions. For continuous outcomes, 53% (17/32) failed to report minimally relevant differences, and 72% (23/32) did not report variances. These elements are crucial for TSA reproducibility. Moreover, the reproducibility of TSAs was associated with journal impact factors and adherence to the PRISMA guidelines. A collective effort is needed from systematic review authors, peer reviewers, and journal editors to improve the reproducibility of TSAs.

Keywords Meta-analysis, Reproducibility, Systematic review, Trial sequential analysis

Introduction

Trial sequential analysis (TSA) has been an increasingly used tool to assess the conclusiveness of evidence synthesized from systematic reviews and meta-analyses (SRMAs) [1–3]. TSA incorporates the concept of cumulative meta-analyses, where each study is added to the evidence synthesis sequentially according to its publication time. Due to multiplicity issues arising from multiple hypothesis testing each time a study is added, TSA applies statistically rigorous methods to adjust the overall type I and type II error rates, thus reducing the likelihood of false positive and false negative conclusions. Moreover, TSAs can estimate required information sizes (RIS), akin to sample size calculations in clinical trials, which helps to determine whether a meta-analysis has adequate statistical power [4]. If the RIS is not achieved, TSA provides decision boundaries that can help assess the statistical significance (monitoring boundaries) or futility (futility boundaries) of an experimental intervention, in a similar manner to interim analyses of clinical trials. Hereafter, we will refer collectively to monitoring and futility boundaries as decision boundaries.

Transparency and reproducibility are essential in validating the conclusions derived from TSAs [5]. Recent years have marked significant improvements in the

*Correspondence:
Lifeng Lin
lifenglin@arizona.edu
[1] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
[2] Menzies Institute for Medical Research, University of Tasmania, Hobart, TAS, Australia
[3] Department of Biostatistics, University of Florida, Gainesville, FL, USA
[4] Evidence-Based Practice Center, Mayo Clinic, Rochester, MN, USA
[5] Department of Epidemiology and Biostatistics, University of Arizona, 1295 N. Martin Ave., Tucson, AZ 85724, USA

reporting quality of SRMAs, due to checklists such as the PRISMA statement [6]. However, the quality of reporting and reproducibility of TSA is unclear. Table 1 outlines three key components of a TSA: the RIS, decision boundaries, and the Z-curve (comprising Z-statistics from cumulative meta-analyses). It also specifies the reporting elements necessary to facilitate the reproduction of TSAs. The aim of this cross-sectional meta-epidemiological study is to assess the reproducibility of TSAs in recent SRMAs.

Methods

Data collection

This study is reported according to standards of reporting meta-epidemiological studies [8]. In September 2023, we identified a convenience sample of SRMAs by searching PubMed for SRMAs that reported the terms "trial sequential analysis" and "meta-analysis" in the title. We restricted our search to the first 100 articles published in 2023, available either in print or online. We excluded protocols and articles that were not written in English.

The rationale for selecting these articles was based on several considerations. First, a large volume of TSA publications has emerged in recent years, particularly following the COVID-19 pandemic [9]. By focusing on a smaller subset of studies from 2023, we were able to manage and analyze the data more effectively and thoroughly. This study was not intended to provide an exhaustive review of all TSA publications; rather, our goal was to highlight issues related to reproducibility using a representative sample. Additionally, limiting the search to recent publications in 2023 ensures the study reflects current practices in TSA. By selecting the first 100 articles without imposing further constraints, such as on journals or research topics, we aimed to minimize potential bias to the best of our ability.

Data extraction

We extracted the TSAs conducted for the primary outcome in each SRMA, including both continuous and binary outcomes. The following data were extracted to reproduce the RIS in TSAs. For continuous outcomes, we collected the type I and type II error rates, diversity (used to adjust between-study heterogeneity in the calculation of RIS), minimally relevant differences, and variances of the continuous outcomes. For binary outcomes, we extracted type I and type II error rates, diversity, relative risk reductions, and assumed event rates in control groups. To reproduce decision boundaries and Z-curves (formed by the Z-statistics in meta-analyses) in TSAs, we also extracted publication years of individual studies, the sample size, mean and standard deviations (continuous outcomes), and event counts and sample sizes (binary outcomes). Notably, all extracted data were directly obtained from the original publications. In this study, we did not evaluate the appropriateness of the input parameters, such as type I and type II error rates, or the data used for TSAs.

Reproducibility

We reproduced TSAs using the TSA 0.9.5.10 Beta software, which is the most commonly utilized tool [10]. For TSAs that did not employ this software, we utilized the

Table 1 Checklist for reporting methods used for performing TSAs

Element in TSA	Reporting item
RIS	• Type I error rate • Type II error rate (or statistical power) • Diversity (if heterogeneity is present) • Minimally relevant differences and variances for continuous outcomes • Relative risk reductions and assumed event rates in control groups for binary outcomes
Decision boundaries	• Data used for deriving information fractions (typically the cumulative sample sizes of individual studies divided by the RIS) • Spending functions for deriving adjusted type I and type II error rates for decision boundaries (optional, as they are typically used as the functions suggested by Lan and DeMets [7])
Z-curve	• Sample means, sample standard deviations, and sample sizes from individual studies for continuous outcomes • 2 × 2 tables (event counts and sample sizes) from individual studies for binary outcomes • Meta-analytical model types, such as the common-effect model (also known as the fixed-effect model) and random-effects model • Estimation methods, particularly for between-study variances, such as the DerSimonian–Laird approach or restricted maximum-likelihood approach • Methods for handling zero events, such as continuity correction, removal, and use of exact models (e.g., generalized linear mixed models)

"metacumbounds" command in Stata 18 and R (version 4.2.1) [11].

Figure 1 illustrates the flowchart for our reproducibility analysis. We categorized the collected articles into three tiers with six distinct groups. In the first tier, SRMAs lacking essential information for computing RIS were placed in group A, while those containing such information were placed in group B.

In the second tier, articles that failed to provide data necessary for deriving decision boundaries and Z-curves were classified as group C; those providing such information were classified as group D. In this tier, we did not further classify based on decision boundaries and Z-curves separately. This decision was primarily because both elements typically require similar meta-analysis data in practice. For example, sample sizes are used to derive information fractions, which in turn yield decision boundaries, while effect sizes and standard errors produce Z-statistics, which form the Z-curve. If the data for reproducing one element (decision boundaries or Z-curve) are unavailable, it is likely that the data for reproducing the other will also be unavailable.

Finally, in the third tier, articles from group D were further divided into group E (where main TSAs could not be reproduced) and group F (where main TSAs could be successfully reproduced). Group F also included all TSAs from groups A and C.

Of note, some TSAs in group F were missing specific details necessary for full reproducibility, such as the meta-analysis model type or the zero-event correction method. We endeavored to reproduce the TSA using the options available in the software. If TSAs could be reproduced through these attempts, we classified the corresponding articles into group F.

Furthermore, we examined the relationship between the reproducibility of TSAs and compliance with the PRISMA statement, as well as journal characteristics, including publication in mega-journals [12] and journal impact factors (IFs), by comparing the proportion of reproducible TSAs among different subgroups. Here, mega-journals refer to academic journals characterized by a large-scale publishing model, broad disciplinary scope, high publication volume, and an emphasis on methodological soundness rather than perceived novelty or impact.

Results

Our final sample consisted of 98 SRMAs (2 excluded due to being published in languages other than English). Figure 2 illustrates the selection process. Complete information about the articles is available at https://osf.io/sngtq/. Among these, 32 TSAs dealt with continuous outcomes and 72 with binary outcomes. All TSAs, except for two, initially utilized TSA software.

Figure 3 presents summaries of missing information for reproducing RIS. For both types of outcomes, the nominal value of the type I error rate was reported in 91% (89/98) of SRMAs, while the nominal value of the type II error rate appeared in 90% (88/98). Only 13% (13/98) of SRMAs reported diversity. Certain TSAs applied the "model variance based" setting within the TSA software but did not explicitly report diversity. Among the 72 binary outcome TSAs, 65% (47/72) failed to report event rates in control groups, and 44% (32/72) did not report relative risk reductions. For zero-event scenarios, only 3 of 69 TSAs involving zero events reported the continuity correction method for zero events. For the 32 continuous outcome TSAs, 53% (17/32) did not report the minimally relevant differences, and variance information was missing in 72% (23/32).

Figure 4 and Table 2 present the proportions of TSAs based on the tiers of reproducibility as classified in

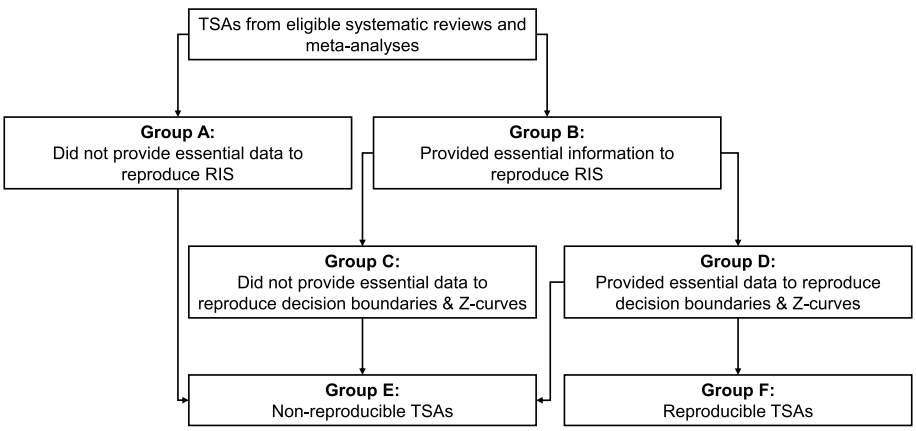

Fig. 1 The flowchart of the TSA reproduction process

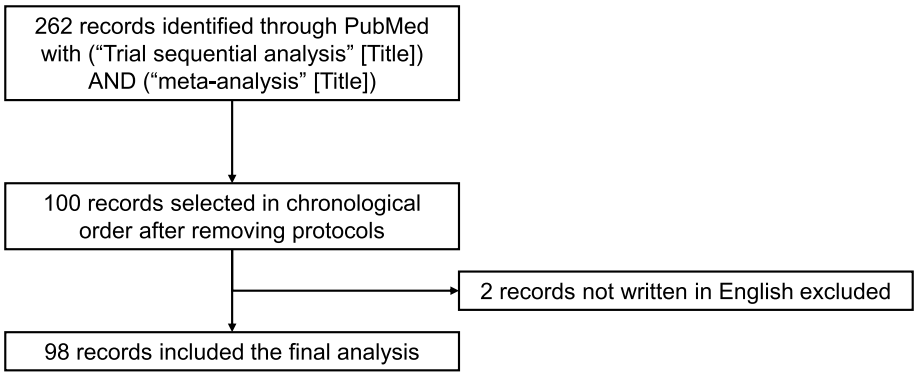

Fig. 2 The flowchart of the data selection process

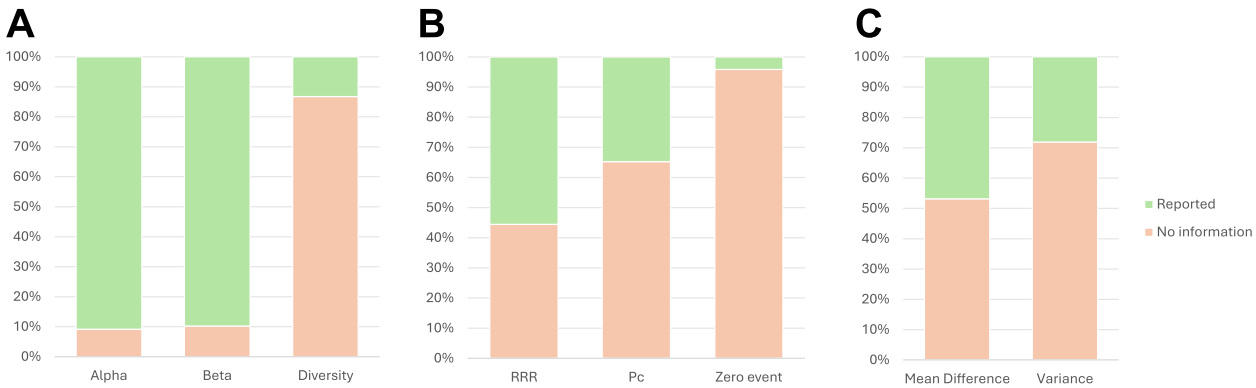

Fig. 3 The charts of studies with missing information for reproducing RIS regarding **A** essential data for both outcomes, **B** specific data for binary outcomes, and **C** specific data for continuous outcomes. Alpha, type I error rate; Beta, type II error rate; RRR, relative risk reduction; Pc, event rate in the control group; Zero event, correction method for zero events

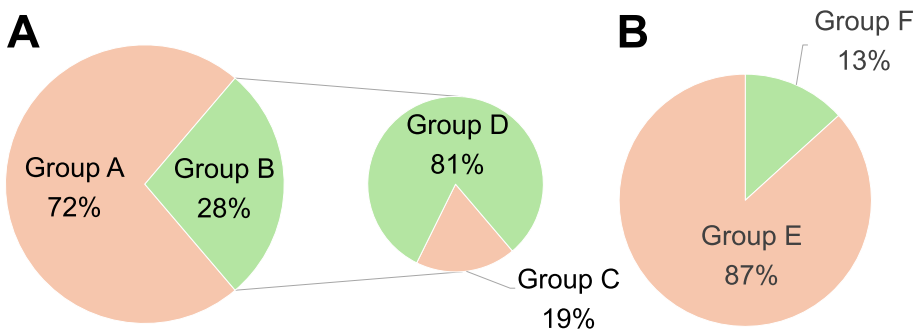

Fig. 4 The pie charts of studies with different reproducible levels **A** for groups A, B, C, and D and **B** for groups E and F. Group A: TSAs that did not provide essential data to reproduce RIS; group B: TSAs that provided essential information to reproduce RIS; group C: TSAs that did not provide essential data to reproduce decision boundaries and Z-curves; group D: TSAs that provided essential data to reproduce decision boundaries and Z-curves; group E: non-reproducible TSAs; group F: reproducible TSAs

Fig. 1. In the first tier, we assessed whether the meta-analyses contained the essential data required to reproduce the RIS. Among the 98 studies evaluated, 27 (28%) contained such essential information (group B), while the remaining 71 meta-analyses lacked this information (group A).

In the second tier, we reviewed TSAs to check for missing data necessary to reproduce decision boundaries and Z-curves. Based on the meta-analyses with the essential data required to reproduce the RIS (group B), 72% (22/27) of these TSAs provided sufficient data in the main texts or supplementary files (group D). Among

Table 2 Proportions of TSAs across different tiers of reproducibility assessment

Tier	Reported information	Group	Proportion
First tier	Essential data for reproducing RIS	Group A (without information)	28% (27/98)
		Group B (with information)	72% (71/98)
Second tier	Essential data for reproducing decision boundaries and Z-curves	Group C (without information)	28% (5/27)
		Group D (with information)	72% (22/27)
Third tier	All essential data for reproducing TSAs	Group E (non-reproducible TSAs)	87% (85/98)
		Group F (reproducible TSAs)	13% (13/98)

the remaining 5 TSAs without such sufficient information (group C), 4 lacked meta-analysis data, and 1 did not include publication years of individual studies.

Finally, in the third tier, we explored the overall reproducibility of TSAs with the essential data. We re-analyzed 22 SRMAs to reproduce TSAs; 13 of these were successfully reproduced (group F), while 9 could not be reproduced (group E). Among these, 11 evaluated binary outcomes and 2 continuous outcomes. Only 4 samples in group F of reproducible TSAs disclosed the models used for the meta-analyses, and only 7 explicitly reported diversity. In the 11 TSAs with binary outcomes, only 2 outlined methods for zero-event correction. The TSA software allowed selections from "Constant," "Reciprocal," and "Empirical" options in the Method drop-down list and values such as 1.0, 0.5, 0.1, and 0.01 in the Value drop-down list; this complicated TSA reproduction without specific information. Ultimately, the main TSAs of 87% (85/98) of SRMAs could not be reproduced, with 13% (13/98) also displaying issues with transparency.

Additionally, none of the TSAs in SRMAs without using the PRISMA statement (0%; 0/20) could be reproduced, while TSAs in SRMAs adhering to the PRISMA statement (17%; 13/78) showed higher reproducibility rates. Among TSAs published in mega-journals, only one (5%; 1/21) could be reproduced; this rate was significantly lower than those not in mega-journals (16%; 12/77). The average journal IF for group F was higher than that for group E (5.55 vs. 5.16). Using a journal IF threshold of 5, 17% (5/30) of TSAs in journals with IF $\geq$ 5 were reproducible, compared to 12% (8/68) in journals with IF < 5.

Discussion

This study examined the reproducibility of a sample of recent TSAs published in 2023. The results highlight a substantial lack of reproducibility of TSAs, mainly stemming from the absence of essential data necessary to compute monitoring and futility boundaries. In this reproducibility study, when information about the meta-analysis model type or the zero-event correction method was missing, we explored various plausible options to reproduce TSAs. Without these efforts, the reproducibility proportion would likely have been even lower. A recent study has documented major errors and inconsistencies in the reporting of TSAs [13], further suggesting a potential reproducibility crisis.

TSAs play a crucial role in providing insights into the conclusiveness of SRMA findings, influencing imprecision judgments and decision-making based on SRMAs. Given this importance, our study identified an urgent need for a standardized guideline outlining the minimal reporting requirements necessary to ensure the transparency and reproducibility of TSAs. Achieving this goal requires a collaborative effort involving systematic review authors, peer reviewers, and journal editors. The brief checklist presented in Table 1 may support this initiative by guiding systematic reviewers in describing the methods used to perform TSAs. We recommend that all essential data and input information for TSAs be included when submitting to journals, ideally as supplementary materials if space is limited. Additionally, systematic reviewers are encouraged to refer to some introductory materials to TSAs, such as Wetterslev et al. [10], Shah and Smith [14], Kang [15], Clephas et al. [2], and Riberholt et al. [16]. These materials provide valuable guidance on conducting TSAs, enhancing understanding of key concepts, and ensuring accurate implementation and interpretation. Such efforts would contribute greatly to improving the reproducibility of TSAs.

There are some limitations to this study. First, it is based on a convenience sample from the first 100 chronological SRMAs from PubMed in 2023. As such, the findings may lack generalizability. We expect that earlier TSAs may even have worse reporting [13]. We acknowledge these limitations and emphasize that the results should be interpreted within the context of these constraints, rather than as a comprehensive representation of all SRMAs using TSAs. Future research could explore reproducibility further by expanding the dataset to include a larger sample of TSAs, such as broadening the database search to multiple sources (e.g., Embase, Cochrane Library) and extending the time frame for study selection. Additionally, employing alternative sampling methods, such as stratified or randomized approaches, could help ensure

the inclusion of studies with diverse characteristics and quality.

Second, this analysis primarily focuses on the statistical data required to reproduce TSAs, yet we did not assess the appropriateness of the input assumptions within these TSAs or their final conclusions. Similar to sample size calculations in clinical trials, TSA conclusions can be influenced by various critical parameters. Therefore, the choice of input values for these parameters requires careful justification and should be a collaborative effort between statisticians and context experts. Future studies that assess TSA conclusions holistically, such as examining the consistency between the conclusions of original studies and re-analyses with appropriate TSA parameters, as well as comparing effect estimates, statistical significance, and clinical relevance, are highly encouraged.

Abbreviations

IF	Impact factor
RIS	Required information size
SRMA	Systematic review and meta-analysis
TSA	Trial sequential analysis

Acknowledgements
We thank the two reviewers for their insightful comments, which have greatly enhanced the quality of our manuscript. ChatGPT 4o was used solely to improve the writing of this manuscript and was not employed for any other purposes in this study.

Authors' contributions
Xing Xing: data extraction, formal analysis, writing—original draft; Yining Wang: data extraction, writing—review and editing; Yipeng Wang: formal analysis, writing—review and editing; Mohammad Hassan Murad: writing—review and editing; Lifeng Lin: supervision, formal analysis, writing—review and editing.

Funding
This study was supported in part by the US National Institute of Mental Health grant R03 MH128727, the US National Library of Medicine grants R21 LM014533 and R01 LM012982, and the Arizona Biomedical Research Centre grant RFGA2023-008–11. The content is solely the responsibility of the authors and does not necessarily represent the official views of the US National Institutes of Health and the Arizona Department of Health Services.

Data availability
The data that support the findings of this research are available at https://osf.io/sngtq/.

Declarations

Ethics approval and consent to participate
Ethics approval and consent to participate were not required for this study, as it used published, de-identifiable data.

Consent for publication
Not applicable.

Competing interests
The authors report there are no competing interests to declare.

Received: 22 March 2024 Accepted: 6 March 2025
Published online: 18 March 2025

References

1. Wetterslev J, Thorlund K, Brok J, Gluud C. Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis. J Clin Epidemiol. 2008;61(1):64–75.
2. Clephas PRD, Kranke P, Heesen M. How to perform and write a trial sequential analysis. Anaesthesia. 2023;78(3):381–4.
3. Murad MH, Wang Z, Chu H, Lin L, El Mikati IK, Khabsa J, Akl EA, Nieuwlaat R, Schuenemann HJ, Riaz IB. Proposed triggers for retiring a living systematic review. BMJ Evid Based Med. 2023;28(5):348–52.
4. Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, Devereaux PJ, Montori VM, Freyschuss B, Vist G et al. GRADE guidelines 6. Rating the quality of evidence—imprecision. J Clin Epidemiol. 2011;64(12):1283–93.
5. Riberholt CG, Olsen MH, Milan JB, Gluud C. Major mistakes and errors in the use of trial sequential analysis in systematic reviews or meta-analyses – protocol for a systematic review. Syst Rev. 2022;11(1):114.
6. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ. 2021;372:n160.
7. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. Biometrika. 1983;70(3):659–63.
8. Murad MH, Wang Z. Guidelines for reporting meta-epidemiological methodology research. BMJ Evid Based Med. 2017;22(4):139–42.
9. Xing X, Wang Y, Lin L. Trial sequential analysis involving same-year studies requires careful temporal ordering. J Clin Epidemiol. 2025;179:111645.
10. Wetterslev J, Jakobsen JC, Gluud C. Trial sequential analysis in systematic reviews with meta-analysis. BMC Med Res Methodol. 2017;17(1):39.
11. Miladinovic B, Hozo I, Djulbegovic B. Trial sequential boundaries for cumulative meta-analyses. Stata J. 2013;13(1):77–91.
12. Ioannidis JPA, Pezzullo AM, Boccia S. The rapid growth of mega-journals: threats and opportunities. JAMA. 2023;329(15):1253–4.
13. Riberholt CG, Olsen MH, Milan JB, Hafliðadóttir SH, Svanholm JH, Pedersen EB, Lew CCH, Asante MA, Pereira Ribeiro J, Wagner V, et al. Major mistakes or errors in the use of trial sequential analysis in systematic reviews or meta-analyses – the METSA systematic review. BMC Med Res Methodol. 2024;24(1):196.
14. Shah A, Smith AF. Trial sequential analysis: adding a new dimension to meta-analysis. Anaesthesia. 2020;75(1):15–20.
15. Kang H. Trial sequential analysis: novel approach for meta-analysis. Anesth Pain Med. 2021;16(2):138–50.
16. Riberholt CG, Olsen MH, Gluud C. Research note: trial sequential analysis in systematic reviews with meta-analysis. J Physiother. 2024;70(3):243–6.

Publisher's Note