International Journal of Epidemiology, 2025, **54(4)**, dyaf098
https://doi.org/10.1093/ije/dyaf098
Opinion

IEA
International Epidemiological Association
OXFORD

Opinion

Multiplicity corrections in life sciences: challenges and consequences

Otília Menyhárt[1,2,3], and Balázs Győrffy[1,3,4,]*

[1]Department of Bioinformatics, Semmelweis University, H-1094 Budapest, Hungary
[2]Cancer Biomarker Research Group, Institute of Molecular Life Sciences, Research Centre for Natural Sciences, H-1117 Budapest, Hungary
[3]National Laboratory for Drug Research and Development, Research Centre for Natural Sciences, H-1117 Budapest, Hungary
[4]Department of Biophysics, Medical School, University of Pecs, H-7624 Pecs, Hungary

*Corresponding author. Department of Bioinformatics, Semmelweis University, Tűzoltó u. 7-9, H-1094 Budapest, Hungary. E-mail: gyorffy.balazs@yahoo.com.

Introduction

The issue of nonreproducibility has become a pervasive challenge across life sciences, raising concerns about the reliability of research findings. It has been argued that current practices may lead to a substantial proportion of false-positive results in published research findings under certain conditions [1]. This concern regarding inflated error rates and decreased reproducibility, resulting from probability pyramiding, was highlighted decades ago, emphasizing the need for independent replication in research [2]. While psychology and social sciences have faced notable criticism for low reproducibility rates, medical research is not exempt. Highly cited research studies often report substantial initial effects that are contradicted or diminished in follow-up investigations [3]. For instance, the large-scale "Reproducibility Project: Cancer Biology" found consistent results in just 26% of attempted replications, with replication effect sizes averaging 85% smaller than initially reported [4].

Central to the reproducibility crisis is the growing complexity of modern research designs, especially in clinical trials and epidemiology. Trials often feature multiple endpoints, treatment arms, or exploratory analyses, whereas epidemiological studies rely on large datasets with numerous variables, each introducing the potential for multiple comparisons. Without proper statistical adjustments, this multiplicity increases the risk of false positives, thereby undermining the reliability of the findings. These risks are compounded by academia's "publish or perish" culture, which favors novel, significant results over rigorous, reproducible work. Selective reporting, p-hacking, and inadequate multiple-hypothesis adjustments distort the evidence base, waste resources, and can lead to misguided clinical decisions [5].

Multiplicity-related issues are particularly concerning in pharmaceutical development, where late-stage trial failures result in significant financial and time losses [3]. Addressing these challenges requires robust statistical practices, prespecified analytical plans, and appropriate corrections for multiple comparisons. Despite long-standing calls for methodological rigor, the application of these adjustments remains inconsistent, even in high-impact journals.

This Opinion paper examines the prevalence and impact of multiplicity-related errors in clinical and epidemiological research. By identifying key gaps in current practices, we propose actionable solutions to enhance the reproducibility and reliability of scientific findings.

Multiplicity in clinical trials

Multiplicity is an inherent challenge in most randomized clinical trials (RCTs). Sponsors often aim to maximize insights from a single study, leading to complex designs that test multiple doses, regimens, endpoints, treatment arms, predictors, subgroups, or populations, often repeated over time. As John Tukey noted in his 1977 paper, "once multiple questions are to be asked, there will be pressures, some ethical in nature, to concentrate upon those questions for which the results appear most favorable" [6].

While multi-arm trials enhance efficiency by reducing the sample size required compared to separate trials, they also amplify the risk of false positives. For instance, in a trial with multiple comparisons, an unadjusted significance level of $\alpha = 0.05$ can lead to substantial inflation of the type I error—a study testing five hypotheses has approximately a 23% probability of producing at least one false positive. A systematic review of 1351 randomized trials published in PubMed in a single month of 2012 revealed that 79% were parallel-group trials, with 14% involving three arms and 7% having four or more arms [7]. Despite the increasing prevalence of multi-arm RCTs, inadequate correction for multiplicity remains a persistent issue, undermining the validity of reported findings.

Prespecifications of analyses to prevent p-hacking

Addressing multiplicity during the design phase of clinical trials is essential to avoid bias, particularly p-hacking, where

Received: 18 December 2024; **Editorial Decision:** 7 May 2025; **Accepted:** 29 May 2025

investigators selectively adopt analysis strategies based on preliminary data review [8]. Established guidelines such as the International Council for Harmonisation (ICH) E9 guideline, *Statistical Principles for Clinical Trials,* or the SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials) stress the need to prespecify statistical methods but provide limited guidance on how analyses should be conducted. To enhance transparency and prevent p-hacking, Kahan *et al.* [8] proposed the pre-SPEC framework, which includes (i) prespecifying analyses before recruitment, (ii) defining a single primary analysis strategy, (iii) creating detailed plans for each analysis, (iv) providing enough detail for independent replication, and (v) ensuring adaptive strategies follow predetermined decisions. Researchers can ensure rigorous, transparent, and reproducible trial analyses by incorporating these measures.

Confirmatory versus exploratory studies

The distinction between confirmatory and exploratory studies should be established early. Confirmatory studies focus on a small set of predefined primary outcomes, with a statistical analysis plan that accounts for family-wise error rate (FWER), *P*-value adjustments, and transparent reporting [9]. On the other hand, exploratory studies are hypothesis-generating and may not require stringent multiple-testing adjustments provided that findings are clearly presented as preliminary hypotheses to be validated by subsequent confirmatory studies. Unfortunately, exploratory results are frequently presented with stronger claims than warranted, leading to overstated conclusions and elevated type I error risks. Therefore, even in exploratory trials, type I error risks should be explicitly acknowledged, and transparent reporting of FWER is advisable [9].

When to adjust for multiple testing

Multiplicity adjustments are critical in studies with: (i) multiple endpoints (e.g. for example, three measures of cardiovascular outcomes); (ii) repeated measures over time (e.g. at three, six, and 12 months); or (iii) multiple treatment arms (e.g. different regimens compared to a shared control arm) [9]. Adjustments are necessary when findings are interpreted collectively for clinical recommendations. For example, in a trial evaluating different schedules or doses of treatment compared to shared control, each experimental arm is part of a family of comparisons and contributes to a broader question, indicating the necessity for multiplicity adjustments [10]. However, adjustments may not be needed if distinct hypotheses inform separate claims of effectiveness. Additional considerations include trials addressing controversial topics or those involving multiple treatments from the same manufacturer, where uncorrected multiplicity may bias conclusions [10].

Multiple endpoints

In trials with coprimary endpoints, where success depends on demonstrating treatment effects across all outcomes, type I error is not inflated, and multiplicity adjustments are unnecessary. These suggestions are consistent with regulatory guidelines for trials aiming for marketing authorization. However, adjustments are required when trials allow multiple pathways to success—where efficacy in any one outcome is sufficient [11].

Related treatments typically demand adjustments for multiple treatment groups, while unrelated treatments may not [9, 10]. Strategies to mitigate type I errors include prioritizing a single primary outcome, using composite outcomes (though challenging due to varying importance to patients), or applying global tests that sum standardized effect sizes. The win ratio approach offers an alternative by prioritizing critical components, such as cause-specific mortality, and calculating effectiveness through patient-pair comparisons [10, 12].

Adjustments for multiple testing in contemporary clinical trials

Despite established guidelines, implementing multiple-testing adjustments in clinical trials remains suboptimal. A 2012 review of multi-arm trials found that only 62% of studies requiring adjustments accounted for multiplicity, with 38% using ordered comparisons and 18% employing single-step procedures, such as the Bonferroni method. Alarmingly, 15% of trials with planned adjustments failed to report them in final publications, suggesting selective reporting [13]. Our review of RCTs and high-profile publications (2010–22) confirms the inconsistent use of correction methods across disciplines, which contributes to inflated Type I error rates and irreproducible findings (Table 1).

Multiplicity adjustments remain inconsistently applied across many research areas. In specific fields, including neurology and psychiatry, multiple primary outcomes are necessary to evaluate the effectiveness of an intervention, as long-term mental health conditions require more than one outcome to characterize the effects of treatment sufficiently. Of 55 RCTs on depression (2007–8), only 5.8% accounted for multiplicity, despite most trials reporting multiple primary and secondary outcomes [14]. Similarly, in a review of 209 neurology and psychiatry RCTs (2011–14), 29% involved multiple primary outcomes, yet 75% lacked adjustments; the Bonferroni correction was the most common method when applied [15]. The review was based on clinical trials published in high-impact journals (*The New England Journal of Medicine, The Lancet, The American Journal of Psychiatry, JAMA Psychiatry, The Lancet Neurology*, and *Neurology*) and had multiplicity has been addressed, some studies would not remain significant [15].

Trials in pain research frequently neglect adjustments. A 2014 review of 161 RCTs revealed that only 52% specified a primary analysis, and 45% applied corrections. None of the articles that neglected to adjust for multiple analyses acknowledged it in their Methods, Results, or Discussion sections. The 15 studies that reported statistical adjustments used mainly the Holm ($n = 3$ studies), Bonferroni ($n = 2$), and Hochberg ($n = 2$) methods [16]. A systemic review of non-pharmacological pain trials with multiple comparisons found that only 21% of RCTs employed statistical adjustments, primarily using the Bonferroni correction [17].

A systematic review of 388 surgical RCTs (2008–20) identified multiplicity in 175 trials. Adjustments were performed in only 20%, primarily using Bonferroni and Tukey methods. Reporting bias affected 51.7% of studies, undermining reliability [18].

Multiplicity is especially critical in imaging, where studies test dozens of hypotheses. A systematic review of PET/CT

Table 1. Reviews investigating the prevalence of adjustments for multiple comparisons in clinical and epidemiological studies

Study (first author)	Year published	Study interval	Scientific field	Number of studies investigated	Number of studies with multiple comparisons	Proportion of studies with adjustments for multiple testing	Most common method	Journals
Wason *et al.*	2014	2012	Multiarm clinical trials	59		51%	Hierarchical/closed and Bonferroni	*British Medical Journal*, *The Lancet*, *New England Journal of Medicine*, and *PLoS Medicine*
Tyler *et al.*	2011	January 2007 and October 2008	Neurology and psychiatry, clinical trials	55	nearly half (52 reported ≥2 primary or secondary outcomes)	3/52 = 5.8%	Bonferroni	Six medical journals
Vickerstaff *et al.*	2015	July 2011–June 2014	Neurology and psychiatry	209	60 (29% reported multiple primary outcomes)	15/60 = 25%	Bonferroni (used in 6 of 15 adjusted trials)	*The New England Journal of Medicine*, *The Lancet*, *The American Journal of Psychiatry*, *JAMA Psychiatry*, *The Lancet Neurology*, and *Neurology*
Kirkham *et al.*	2015	2012	Otolaryngology, original human subjects research with samples of >100 subjects	195	140	10%	Bonferroni (used in 5 of 8 corrected studies)	*The Laryngoscope*; *Archives of Otolaryngology—Head & Neck Surgery* (now called *JAMA Otolaryngology—Head & Neck Surgery*); *Otolaryngology—Head & Neck Surgery*; and *Annals of Otology, Rhinology, and Laryngology*.
Stacey *et al.*	2012	2010	Ophthalmology, abstracts presented at ARVO 2010	6415 abstracts	538 (with more than five hypotheses)	14%	Bonferroni and Tukey	
Dworkin *et al.*	2016	January 2006 and June 2013	Pain, randomized clinical trial	101	29	21%	Bonferroni, gatekeeping, Sidák	*European Journal of Pain*, the *Journal of Pain*, and *Pain*.
Gewandter	2014	2006 and 2012	Pain, randomized clinical trial	160	33	45%		*European Journal of Pain*, the *Journal of Pain*, and *Pain*.
Chalkidou	2014	2000–13	PET CT scan	15	15	1 study	Unspecified	
Brand	2021	2015–16	Cardiovascular, randomized clinical trial	130	89 studies with subgroups	2 studies altogether	Unspecified	*The New England Journal of Medicine*, *The Lancet* and *Circulation*.
Nevins	2022	2014–19	Pragmatic clinical trials	262 final reports, 159 protocols	38 final reports, 30 protocols	11% final reports, 7% protocols	Bonferroni	MEDLINE

(continued)

Table 1. (continued)

Study (first author)	Year published	Study interval	Scientific field	Number of studies investigated	Number of studies with multiple comparisons	Proportion of studies with adjustments for multiple testing	Most common method	Journals
Pike	2022	January to June 2018	Randomized clinical trial, general medicine	138	28	48% for multiple treatment comparisons (11 out of 28)	Bonferroni, Holm, Hochberg, hierarchical methods	7 journals: *Annals of Internal Medicine*, *British Medical Journal (BMJ)*, *Journal of the American Medical Association (JAMA)*, *Lancet*, *New England Journal of Medicine (NEJM)*, *National Health Technology Assessment (HTA)*, or *PLOS Medicine (PLoS Med)*
Wallach	2018	SATIRE (2007), DISCO (2002–12)	Randomized clinical trial (various medical fields (e. g. cardiovascular, infectious disease)	64 (with at least one subgroup claim in the abstract)	46 (subgroup findings with statistical support)	1/46 =2.2%	Bonferroni–Holm step-down procedure	
Benjamini	2010	2000–4	Articles, samples from *NEJM*	60	All had multiplicity issues	22%	Minimal, mostly no methods used	*New England Journal of Medicine*
Benjamini and Cohen	2017	2000–10	Articles, samples from *NEJM*	100	All had multiple endpoints	Only 20% addressed multiplicity in any form	Very few used any method	*New England Journal of Medicine*

biomarker studies (2000–13) found 15 eligible studies, with the number of hypotheses ranging between 8 and 102. Only one study applied adjustments for multiple hypothesis testing, and three studies included validation of their results. The authors estimated an average type-I error probability of 76% (34%–99%), with most published results failing to reach statistical significance [19]. Similarly, unadjusted *P*-values remain an issue in fMRI research: in the Brede database of neuroimaging papers, 49% of studies ($n = 1705$ articles) published uncorrected *P*-values [20].

In ophthalmology, only 14% of abstracts at a major conference with more than five hypothesis tests reported corrections [21]. Similarly, a systematic review of otolaryngology studies found that 72% involved multiple testing, but only 10% addressed it, primarily using Bonferroni corrections. The authors estimated a 41% chance of false positives and noted that 18% of reported *P*-values might be spurious [22].

Multiplicity challenges in pragmatic trials

Pragmatic trials, designed to evaluate real-world effectiveness, often include multiple primary outcomes but rarely apply corrections. Only 10% of such trials (2014–19) reported adjustments. In a review of high-impact journals, inconsistencies were noted: 25% of trials adjusted when all outcomes had to show effectiveness, but only 35% adjusted when not all outcomes needed to succeed. Opinions among statisticians were similarly divided, with adjustments for multiple primary outcomes seen as critical by some but less essential for secondary outcomes or subgroup analyses [23, 24].

Multiplicity challenges in high-impact publications

Even high-profile journals frequently overlook multiplicity adjustments, leading to inflated type I errors. Benjamini and colleagues investigated a batch of papers published in the *New England Journal of Medicine* (NEJM) between 2000 and 2004. Out of the 60 papers, 47 (78.3%) had no multiplicity adjustments, although they should have had in some form [25]. The trend remained, as 80% of papers published between 2000 and 2010 in NEJM investigating multiple endpoints ignored the issue of multiplicity. In multi-arm clinical trials published in 2012 by four top-quality medical journals (*British Medical Journal, The Lancet, New England Journal of Medicine*, and *PLoS Medicine*), 51% of studies contained adjustment for multiplicity, with a slightly higher proportion of exploratory trials (55%) compared to confirmatory studies (46%). The most common adjustment methods were the gatekeeping/hierarchical/closed approach (24%) and the Bonferroni correction (14%) [9]. A study published in 2020 investigating cardiovascular randomized trials in six high-impact journals found that out of 300 studies with multiplicity, only 28% (85 studies) adjusted the results for multiple comparisons. Interestingly, larger trials were less likely to make adjustments [26].

Multiplicity affecting subgroup analyses

The treatment effect of a new intervention may vary among different segments of the study population. Subgroup analyses may assess the safety profiles and the consistency of the treatment effect across subgroups and detect effects within a subgroup in an otherwise nonsignificant trial. Subgroup analyses may be predefined or a posteriori and may be delineated as confirmatory, exploratory, *post hoc*, and data-driven subgroup identifications [27, 28]. Subgroup analyses carry a high risk of false positives, especially when numerous comparisons are made. A review of 89 cardiovascular studies showed more common subgroup analyses in trials with nonsignificant primary results, highlighting potential "fishing for significance." Only 2% of these trials applied adjustments [29]. A meta-analysis of 64 RCTs with 117 subgroup claims found that only one study used the Bonferroni–Holm procedure, while 39.3% lacked statistical evidence for their claims [30]. Guidelines recommend corrections in subgroup-based comparisons but suggest flexibility when subgroups are nested within treatment arms [31].

Recommended adjustment methods

Effective control of multiplicity is essential to ensure the validity of clinical trial results and prevent erroneous conclusions about treatment efficacy. Adjustment methods must align with the specific clinical and statistical context, considering endpoint priorities, population differences, and correlations between hypotheses. Selecting the optimal approach requires extensive trial simulations to balance sensitivity and specificity.

Multiplicity adjustment strategies in clinical trials generally fall into two extremes: "no adjustment at all" or excessively strict "no error at all" approaches (Table 1). The Bonferroni correction remains the most widely used method for controlling the family-wise error rate (FWER). However, it is often criticized for being overly conservative, increasing the risk of false negatives, and potentially overlooking meaningful effects. Stepwise methods, such as Holm's procedure (adjusting significance thresholds sequentially) and Hochberg's method (ranking hypotheses to set critical values), offer a more balanced approach by controlling type I error rates while maintaining statistical power. These methods are increasingly preferred for their ability to balance stringency and flexibility [32].

False discovery rate (FDR) adjustments provide an alternative approach for large-scale studies involving numerous hypotheses, such as genome-wide association studies. FDR-based methods control the proportion of false positives among significant results, yielding q-values that balance sensitivity and specificity. These approaches are highly effective for preserving statistical power while minimizing false discoveries [33]. Our prior publication details frequently used adjustment methods, including FWER and FDR-based techniques [32].

Moreover, accessible tools for multiplicity correction offer researchers practical means to maintain statistical rigor. One freely available example is "multipletesting.com," developed by the authors [32]. This intuitive tool requires no coding skills or specialized expertise, making a multiplicity correction broadly accessible to many users. Additional resources include publicly available FDR adjustment calculators and standard software functions, such as the p.adjust function in R.

Gatekeeping procedures handle multiplicity in scenarios where hypotheses are hierarchically or logically related. Primary endpoints are tested first, with secondary endpoints evaluated only if primary tests are significant, safeguarding statistical integrity across multiple comparisons. Methods such as Bonferroni and Holm are common in gatekeeping frameworks, with variants like serial, parallel, and tree-structured approaches tailored for complex trial designs.

Clinical examples show how gatekeeping methods address specific trial design challenges [34, 35].

In drug development, weighted FDR methods account for the relative importance of different endpoints. They achieve greater power by prioritizing primary endpoints and assigning weights to secondary outcomes than traditional hierarchical gatekeeping. These weighted procedures are beneficial in phase II trials where primary endpoints may have low power, enhancing secondary outcome discovery while maintaining rigorous multiplicity control. Such methods apply across diverse research fields where hypotheses vary in importance [36].

Multiplicity in observational epidemiology

Observational epidemiology, which relies on biomarkers or questionnaires to assess exposures, struggles with reproducibility, compounded by high-throughput technologies like transcriptomics and metabolomics [37]. The exposome concept, encompassing all environmental and biological exposures, is vital to understanding gene–environment interactions. Nevertheless, the number of variables in exposome studies often leads to high type I error rates if multiplicity is not addressed [38].

Debate persists over whether multiplicity adjustments are needed in observational studies. Rothman [39] argued against routine adjustments, warning that they could mask important findings by increasing type II errors—particularly in exploratory contexts. However, Rothman's viewpoint assumes researchers transparently present exploratory findings without overstating their significance, which is frequently not the practice case, as transparency regarding multiplicity is often lacking in publications [38]. While genomic studies (e.g. Genome-wide association studies) routinely employ multiplicity adjustments, this is not the norm in broader epidemiology. A 1998 review of articles in the *American Journal of Epidemiology* and the *American Journal of Public Health* revealed a ∼20% type I error rate, far exceeding the expected 5% [40].

Nutritional epidemiology particularly exemplifies these challenges. Nonrandomized observational studies in this field often employ flexible statistical adjustments and Food Frequency Questionnaires (FFQs), which can lead to questionable validity. This practice has resulted in almost every food ingredient being linked to cancer risk at some point, despite most findings being implausible or irreproducible [41]. Small effect sizes, confounding, measurement errors, and unadjusted multiplicity exacerbate this issue. Critics argue that focusing on single ingredients in observational or small-scale randomized trials will unlikely advance the field [42]. Instead, large-scale trials evaluating dietary patterns, such as the PREDIMED trial on the Mediterranean diet, offer a more reliable approach to understanding complex nutritional impacts [43].

Transparency in observational studies can be enhanced by explicitly reporting every planned comparison, including measured exposures, outcomes, and their inter-relationships. While the actual number of tests performed can be challenging to capture—especially in exploratory contexts—researchers should strive to track all analyses attempted, even those ultimately excluded from the final report [44]. Preregistration of a Statistical Analysis Plan (SAP) further strengthens transparency by clearly distinguishing confirmatory from exploratory hypotheses, mitigating inflated effect sizes, and selective reporting (as well as HARKing—hypothesizing after results are known) [45]. Recent metascientific evidence strongly supports the use of preregistration and Registered Reports, confirming that these practices effectively mitigate systematic and publication bias by enabling peers to critically evaluate how rigorously scientific claims have been tested [46]. Despite occasional criticism—such as preregistration discouraging exploratory research or inadvertently signaling study quality without proper scrutiny—these concerns lack strong empirical backing and often arise from misunderstandings about the goal of preregistration [46]. Thus, preregistration addresses Rothman's concerns by transparently delineating exploratory findings without necessarily discouraging exploration itself.

Multiplicity affecting meta-analyses

Systematic meta-analyses effectively summarize accumulated knowledge, aiding clinicians and policymakers. However, multiplicity issues must be addressed explicitly. Detailed prespecification of decision rules for managing multiple groups, time points, and analyses in meta-analysis protocols prevents selective reporting or cherry-picking significant findings from primary studies. Observational studies, often exploratory by nature, are particularly prone to selective inclusion and inflated biases. Unfortunately, observational studies with multiple comparisons, conducted without corrections for multiple hypotheses, will produce meta-analyses with unreliable results [37]. Meta-analyses based on papers with a risk of containing false positives should be treated with care until the credibility of the underlying primary studies is evaluated or further confirmatory studies are conducted.

Conclusions

The reproducibility crisis highlights the urgent need for rigorous statistical practices. Even RCTs, the methodological gold standard, are susceptible to multiplicity issues. Clear prespecification of analytical methods, proper correction for multiplicity, and transparent reporting are essential. Adopting these practices can minimize false-positive findings, enhance reproducibility, and provide a more informed basis for evidence-based clinical and policy decisions.

Acknowledgements

The support of ELIXIR Hungary (www.bioinformatics.hu) is acknowledged.

Author contributions

Both authors contributed to the conception and design of the study. Data collection and analysis were performed by O.M. The first draft of the manuscript was written by O.M., and both authors commented on previous versions of the manuscript. All authors read and approved the final manuscript. O.M. will act as a guarantor of the paper.

Conflict of interest: None declared.

Funding

This project was supported by the National Research, Development, and Innovation Office (PharmaLab, RRF-2.3.1-21-2022-00015). O.M. was supported by the Janos

Bolyai Scholarship of the Hungarian Academy of Sciences and the Hungarian Scientific Research Fund (OTKA FK147194).

Data availability

NA.

Use of artificial intelligence (AI) tools

AI was utilized to enhance grammar during the writing process.

Ethics approval

NA.

References

1. Ioannidis JP. Contradicted and initially stronger effects in highly cited clinical research. *JAMA* 2005;**294**:218–28.
2. Neher A. Probability pyramiding, research error and the need for independent replication. *Psychol Rec* 1967;**17**:257–62.
3. Begley CG, Ellis LM. Raise standards for preclinical cancer research. *Nature* 2012;**483**:531–3.
4. Errington TM, Mathur M, Soderberg CK *et al.* Investigating the replicability of preclinical cancer biology.*eLife* 2021;**10**:e71601.
5. Jager LR, Leek JT. An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics* 2014;**15**:1–12.
6. Tukey JW. Some thoughts on clinical trials, especially problems of multiplicity. *Science* 1977;**198**:679–84.
7. Odutayo A, Emdin CA, Hsiao AJ *et al.* Association between trial registration and positive study findings: cross sectional study (Epidemiological Study of Randomized Trials—ESORT). *BMJ* 2017;**356**:j917.
8. Kahan BC, Forbes G, Cro S. How to design a pre-specified statistical analysis approach to limit p-hacking in clinical trials: the pre-SPEC framework. *BMC Med* 2020;**18**:253.
9. Wason JMS, Stecher L, Mander AP. Correcting for multiple-testing in multi-arm trials: is it necessary and is it done? *Trials* 2014;**15**:364.
10. Li G, Taljaard M, Van den Heuvel ER *et al.* An introduction to multiplicity issues in clinical trials: the what, why, when and how. *Int J Epidemiol* 2017;**46**:746–55.
11. Committee for Human Medicinal Products (CHMP). *Guideline on Multiplicity Issues in Clinical Trials*. London, UK: European Medicines Agency, 2017.
12. Dmitrienko A, D'Agostino RB Sr. Multiplicity considerations in clinical trials. *N Engl J Med* 2018;**378**:2115–22.
13. Odutayo A, Gryaznov D, Copsey B *et al.*; ASPIRE Study Group. Design, analysis and reporting of multi-arm trials and strategies to address multiple testing. *Int J Epidemiol* 2020;**49**:968–78.
14. Tyler KM, Normand S-LT, Horton NJ. The use and abuse of multiple outcomes in randomized controlled depression trials. *Contemp Clin Trials* 2011;**32**:299–304.
15. Vickerstaff V, Ambler G, King M, Nazareth I, Omar RZ. Are multiple primary outcomes analysed appropriately in randomised controlled trials? A review. *Contemp Clin Trials* 2015;**45**:8–12.
16. Gewandter JS, Smith SM, McKeown A *et al.* Reporting of primary analyses and multiplicity adjustment in recent analgesic clinical trials: ACTTION systematic review and recommendations. *Pain* 2014;**155**:461–6.
17. Dworkin JD, McKeown A, Farrar JT *et al.* Deficiencies in reporting of statistical methodology in recent randomized trials of nonpharmacologic pain treatments: ACTTION systematic review. *J Clin Epidemiol* 2016;**72**:56–65.
18. Robinson NB, Fremes S, Hameed I *et al.* Characteristics of randomized clinical trials in surgery from 2008 to 2020: a systematic review. *JAMA Network Open* 2021;**4**:e2114494.
19. Chalkidou A, O'Doherty MJ, Marsden PK. False discovery rates in PET and CT studies with texture features: a systematic review. *PLoS One* 2015;**10**:e0124165.
20. Saxe R, Brett M, Kanwisher N. Divide and conquer: a defense of functional localizers. *NeuroImage* 2006;**30**:1088–96.
21. Stacey AW, Pouly S, Czyz CN. An analysis of the use of multiple comparison corrections in ophthalmology research. *Investig Ophthalmol Vis Sci* 2012;**53**:1830–4.
22. Kirkham EM, Weaver EM. A review of multiple hypothesis testing in otolaryngology literature. *Laryngoscope* 2015;**125**:599–603.
23. Nevins P, Vanderhout S, Carroll K *et al.* Review of pragmatic trials found that multiple primary outcomes are common but so too are discrepancies between protocols and final reports. *J Clin Epidemiol* 2022;**143**:149–58.
24. Pike K, Reeves BC, Rogers CA. Approaches to multiplicity in publicly funded pragmatic randomised controlled trials: a survey of clinical trials units and a rapid review of published trials. *BMC Med Res Methodol* 2022;**22**:39.
25. Benjamini Y. Simultaneous and selective inference: current successes and future challenges. *Biom J* 2010;**52**:708–21.
26. Khan MS, Khan MS, Ansari ZN *et al.* Prevalence of multiplicity and appropriate adjustments among cardiovascular randomized clinical trials published in major medical journals. *JAMA Netw Open* 2020;**3**:e203082-e.
27. Bunouf P, Groc M, Dmitrienko A, Lipkovich I. Data-driven subgroup identification in confirmatory clinical trials. *Ther Innov Regul Sci* 2022;**56**:65–75.
28. Lipkovich I, Dmitrienko A, B R. Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Stat Med* 2017;**36**:136–96.
29. Brand KJ, Hapfelmeier A, Haller B. A systematic review of subgroup analyses in randomised clinical trials in cardiovascular disease. *Clin Trials (London, England)* 2021;**18**:351–60.
30. Wallach JD, Sullivan PG, Trepanowski JF, Sainani KL, Steyerberg EW, Ioannidis JP. Evaluation of evidence of statistical support and corroboration of subgroup claims in randomized clinical trials. *JAMA Intern Med* 2017;**177**:554–60.
31. Stallard N, Todd S, Parashar D, Kimani PK, Renfro LA. On the need to adjust for multiplicity in confirmatory clinical trials with master protocols. *Ann Oncol* 2019;**30**:506–9.
32. Menyhart O, Weltz B, Győrffy B. MultipleTesting.com: a tool for life science researchers for multiple hypothesis testing correction. *PLoS One* 2021;**16**:e0245824.
33. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B (Methodological)* 1995;**57**:289–300.
34. Dmitrienko A, Wiens BL, Tamhane AC, Wang X. Tree-structured gatekeeping tests in clinical trials with hierarchically ordered multiple objectives. *Stat Med* 2007;**26**:2465–78.
35. Kordzakhia G, Brechenmacher T, Ishida E, Dmitrienko A, Zheng WW, Li DF. An enhanced mixture method for constructing gatekeeping procedures in clinical trials. *J Biopharm Stat* 2018;**28**:113–28.
36. Benjamini Y, Cohen R. Weighted false discovery rate controlling procedures for clinical trials. *Biostatistics* 2017;**18**:91–104.
37. Peace K, Yin J, Rochani H *et al.* The Reliability of a Nutritional Meta-Analysis Study. arxiv, https://arxiv.org/abs/1710.02219, 2017, preprint: not peer reviewed.
38. Patel CJ, Ioannidis JP. Placing epidemiological results in the context of multiplicity and typical correlations of exposures. *J Epidemiol Commun Health* 2014;**68**:1096–100.
39. Rothman KJ. No adjustments are needed for multiple comparisons. *Epidemiology (Cambridge, Mass)* 1990;**1**:43–6.
40. Ottenbacher KJ. Quantitative evaluation of multiplicity in epidemiology and public health research. *Am J Epidemiol* 1998;**147**:615–9.

41. Schoenfeld JD, Ioannidis JP. Is everything we eat associated with cancer? A systematic cookbook review. *Am J Clin Nutr* 2013; **97**:127–34.
42. Ioannidis JPA. Implausible results in human nutrition research. *BMJ* 2013;**347**:f6698.
43. Estruch R, Ros E, Salas-Salvadó J *et al.* PREDIMED Study Investigators. Primary prevention of cardiovascular disease with a mediterranean diet supplemented with extra-virgin olive oil or nuts. *N Engl J Med* 2018;**378**:e34.
44. Dal-Ré R, Ioannidis JP, Bracken MB *et al.* Making prospective registration of observational research a reality. *Sci Transl Med* 2014; **6**:224cm1.
45. Hiemstra B, Keus F, Wetterslev J, Gluud C, van der Horst ICC. DEBATE-statistical analysis plans for observational studies. *BMC Med Res Methodol* 2019;**19**:233.
46. Lakens D, Mesquida C, Rasti S, Ditroilo M. The benefits of preregistration and registered reports. *Evid Based Toxicol* 2024; **2**:2376046.

International Journal of Epidemiology, 2025, 54, 1–8
https://doi.org/10.1093/ije/dyaf098
Opinion