available at www.sciencedirect.com
journal homepage: www.europeanurology.com

European Association of Urology

Platinum Opinion

How To Interpret Subgroup Analyses from Prospective Randomized Clinical Trials: What Clinicians Need To Know

Susan Halabi [a,b,*], *Siyuan Guo* [a]

[a] *Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, NC, USA;* [b] *Duke Cancer Institute Center for Prostate and Urologic Cancers, Durham, NC, USA*

1. Introduction

Subgroup analyses are pervasive among randomized controlled trials (RCTs) in efforts to assess whether treatment effects vary across patient characteristics such as biomarker status, disease burden, or prior therapy. Regardless of the overall trial result, subgroup analyses are routinely conducted. In positive trials, they help to identify who benefits most or least. In negative trials, they often serve as exploratory tools to uncover signals that could inform future clinical trials. While these analyses can generate valuable clinical insights, they come with substantial methodological risks. Too often, noise rather than any biological signal drives the narrative. Key concerns include inflated type I errors, limited power, and misinterpretation of *p* values, which can lead to false-negative and false-positive results.

Most RCTs are designed to detect an average treatment effect, assuming homogeneity of the response across the population. Yet in reality, patient populations are biologically and clinically heterogeneous. Subgroup analyses attempt to capture this variability, but must be conducted and interpreted with rigor and caution to ensure valid and meaningful conclusions.

2. The statistical challenges

2.1. Inflated type I errors

Most RCTs are designed to test the overall treatment effect and not for detection of effects in subgroups. With each additional subgroup comparison, the likelihood of a false-positive result increases. For example, testing five independent subgroups yields a 23% chance of observing at least one *p* value <0.05 by chance alone, which rises to 40% with ten independent subgroups (Fig. 1A). Moreover, the probability of seeing a reversal of a treatment effect in at least one subgroup is surprisingly high. With eight subgroups and a modest treatment effect (standardized difference = 0.2), there is an 89% chance of observing a subgroup in which the direction of an effect appears to be reversed, purely due to chance.

2.2. Lower statistical power

If a subgroup is of interest, the hypothesis should ideally be specified at the planning stage, with control for type I errors and adequate power. For example, the phase 3 AMBASSADOR trial (NCT03244384) was designed with disease-free survival and overall survival (OS) as joint primary endpoints for assessment of the treatment effect in both the overall population and subgroups defined by PD-L1 expression status [1]. The trial maintained rigorous control of type I error at a one-sided level of 0.025 and used an iterative testing procedure [2].

Subgroup analyses often face lower power because of smaller sample sizes and fewer events. For instance, if only 50% of trial events occur in a subgroup, the power for detection of a hazard ratio of 0.75 can decline substantially, even if the study has 85% power (Fig. 1B). This low power increases the risk of false-negative results and instability in estimating the treatment effect in the subgroup.

* Corresponding author. Duke University Medical Center, 2424 Erwin Road, Durham, NC 27710, USA. Tel. +1 919 681 5430; Fax: +1 919 668 7061.
E-mail address: susan.halabi@duke.edu (S. Halabi).

https://doi.org/10.1016/j.eururo.2025.06.015

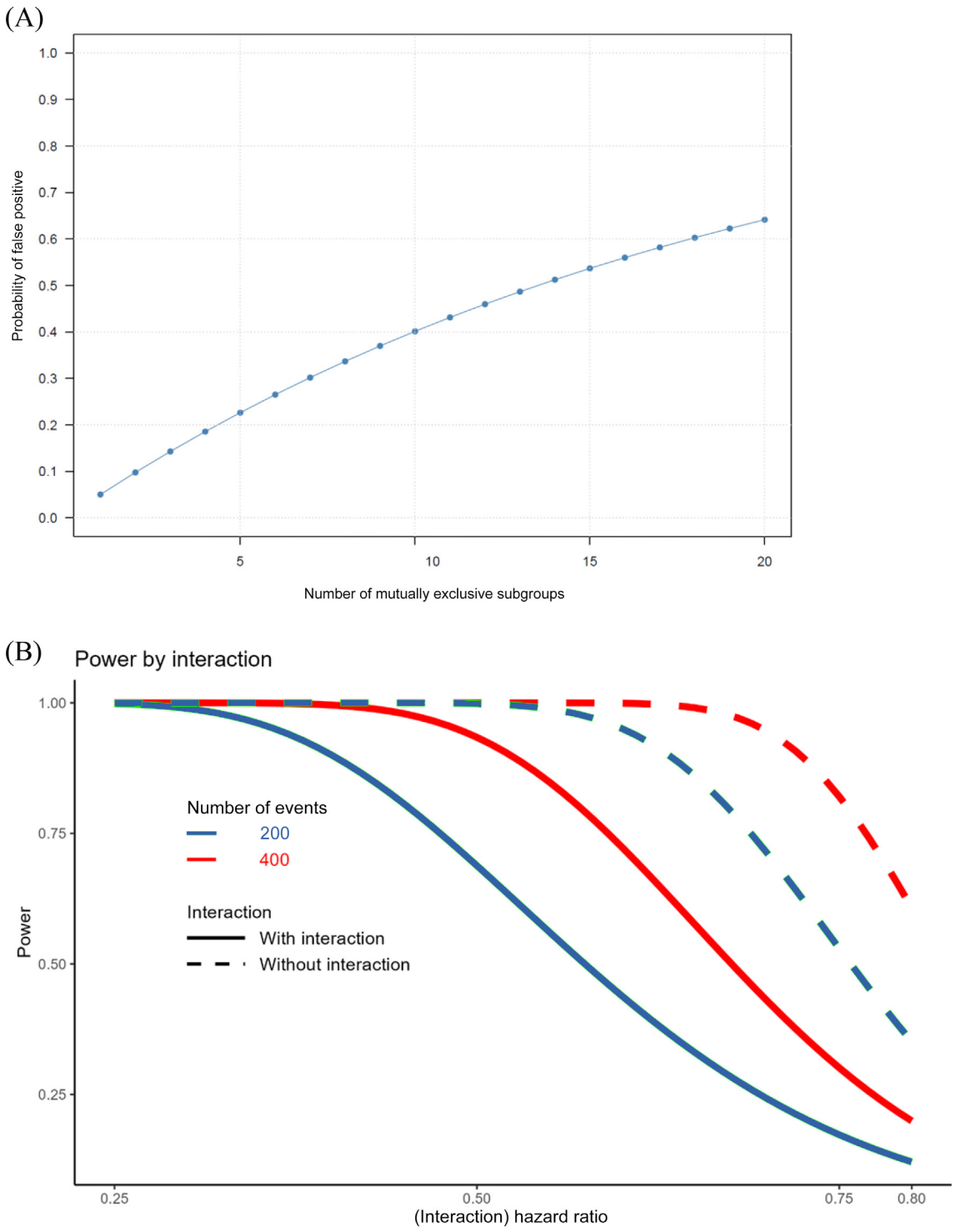

Fig. 1 – (A) Probability of false positives as a function of the number of analyses of mutually independent subgroups. (B) Impact of an interaction effect and sample size on statistical power.

2.3. Visual fallacy: separate subgroup significance tests can be misleading

A common pitfall in subgroup analyses is interpretation of separate tests of the treatment effect within subgroups as evidence of differential efficacy. Suppose a treatment effect is significant overall (eg, $p = 0.045$) and also appears to be significant in one subgroup ($p = 0.034$) but not in another ($p = 0.48$). This does not mean that the treatment works in one group but not the other. Such conclusions are misleading without formally testing of the interaction between treatment and subgroup. The appropriate metric is the interaction hazard ratio (HR), defined as the ratio of HRs across subgroups, which is a direct measure of whether

the treatment effects differ. As shown in Fig. 1B, the statistical power decreases when a true interaction exists, which underscores the importance of planning for such analyses during trial design when supported by a biological rationale [3]. In addition, subgroup misclassification (eg, because of biomarker error) can further reduce the statistical power, so larger sample sizes are needed to adequately detect interaction effects [4].

3. Levels of evidence and the role of meta-analysis

The credibility of a subgroup claim depends on the study design and context. Prespecified analyses with adequate power, significant interaction tests, and a biological rationale carry more weight than post hoc findings (Fig. 2A). In the CHAARTED trial, a prespecified subgroup analysis showed that men with high-disease volume metastatic hormone-sensitive prostate cancer (mHSPC) had longer OS with androgen deprivation therapy (ADT) + docetaxel than with ADT alone (hazard ratio 0.60, 95% confidence interval 0.45–0.81) [5], although the study was not designed to test for an interaction between treatment and disease volume. This was later confirmed in a meta-analysis of three mHSPC trials [6]. Similarly, the STAMPEDE trial found no OS benefit from prostate radiotherapy overall, but a prespecified subgroup analysis suggested benefit for men with low metastatic burden [7], which was confirmed in a pooled meta-analysis [8]. Because single trials are often underpowered for detection of interactions, meta-analyses can help in improving the statistical power, confirming reproducibility, and supporting robust conclusions, and can potentially shape treatment guidelines.

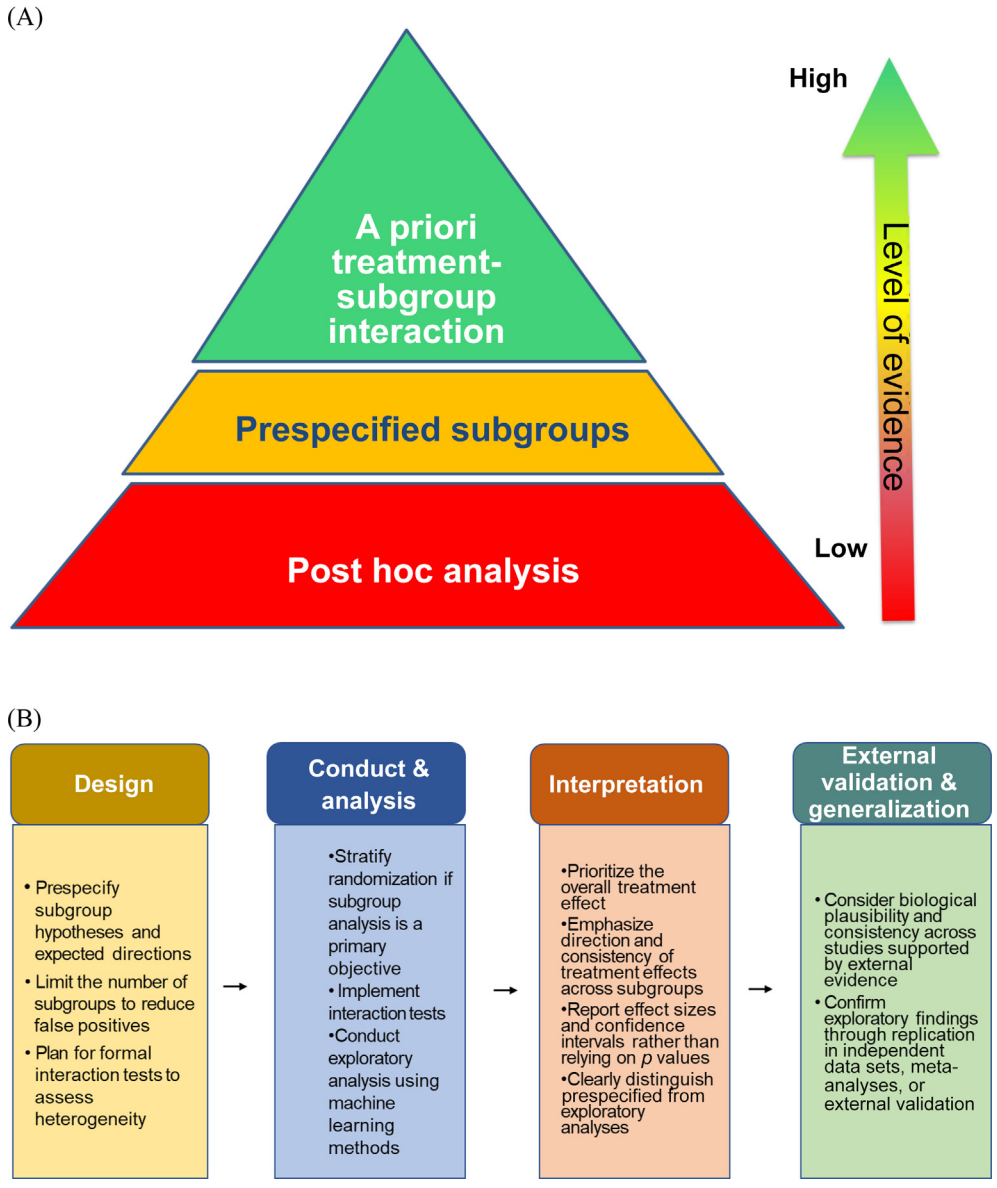

Fig. 2 – Subgroup analysis: (A) hierarchy of evidence and (B) best practices.

4. Traditional versus data-driven subgroup definitions

Subgroup definitions have traditionally relied on clinical judgment and biological rationale, such as classification of patients by tumor stage, biomarker expression, or prior treatments. These categories are grounded in existing knowledge and are easier to interpret and communicate. However, they may overlook more complex predictors of treatment heterogeneity.

By contrast, data-driven methods such as recursive partitioning, clustering, and machine-learning algorithms can uncover unrecognized subgroup effects. For example, an approach called subgroup identification via differential effect search (SIDES) uses tree-based algorithms to identify subgroups with enhanced treatment-effect interactions [9]. While such approaches generate novel hypotheses for future trials, they raise concerns about overfitting, reproducibility, and interpretability. Regulatory acceptance will probably require external validation and biological plausibility. Integration of classical and data-driven approaches, anchored in statistical rigor, represents a promising area for future research.

5. Conclusions

Subgroup analyses must be approached with caution. Unless such analyses are prespecified at the design stage, any findings should be considered exploratory and confirmed in future trials (Fig. 2B). Poorly designed subgroup analyses risk misleading interpretation and may erode trust among clinicians, regulators, and the public. Credible subgroup claims require safeguards such as prespecification, adequate power, formal interaction testing, and meta-analysis. Statistical reporting should emphasize effect sizes and confidence intervals, in line with the 2016 guidance from the American Statistical Association on the appropriate use of p values [10].

Regulatory decisions are based not only on statistical significance, but also on biological plausibility, safety, and the totality of evidence. A well-thought, transparent approach is essential to ensure subgroup findings contribute meaningfully to evidence-based decision-making.

Conflicts of interest: Susan Halabi has served as a member of data safety monitoring boards for BMS, Beigene, CG Oncology, Janssen, and Sanofi, and a consultant for AstraZeneca and Pfizer; and has received institutional research funding from the American Society of Clinical Oncology and the Prostate Cancer Foundation. Siyuan Guo has nothing to disclose.

Acknowledgments: The authors thank Chenxi Yu for creating Fig. 1A.

References

[1] Apolo AB, Ballman KV, Sonpavde G, et al. Adjuvant pembrolizumab versus observation in muscle-invasive urothelial carcinoma. N Engl J Med 2025;392:45–55.
[2] Maurer W, Bretz F. Multiple testing in group sequential trials using graphical approaches. Stat Biopharm Res 2013;5:311–20.
[3] Moser BK, Halabi S. Sample Size requirements and study duration for testing main effects and interactions in completely randomized factorial designs when time to event is the outcome. Commun Stat Theory Methods 2015;44:275–85.
[4] Halabi S, Lin CY, Liu A. On the design and the analysis of stratified biomarker trials in the presence of measurement error. Stat Med 2021;40:2783–99.
[5] Sweeney CJ, Chen YH, Carducci M, et al. Chemohormonal therapy in metastatic hormone-sensitive prostate cancer. N Engl J Med 2015;373:737–46.
[6] Vale CL, Fisher DJ, Godolphin PJ, et al. Which patients with metastatic hormone-sensitive prostate cancer benefit from docetaxel: a systematic review and meta-analysis of individual participant data from randomised trials. Lancet Oncol 2023;24:783–97.
[7] Parker CC, James ND, Brawley CD, et al. Radiotherapy to the primary tumour for newly diagnosed, metastatic prostate cancer (STAMPEDE): a randomised controlled phase 3 trial. Lancet 2018;392:2353–66.
[8] Burdett S, Boeve LM, Ingleby FC, et al. Prostate radiotherapy for metastatic hormone-sensitive prostate cancer: a STOPCAP systematic review and meta-analysis. Eur Urol 2019;76:115–24.
[9] Lipkovich I, Dmitrienko A, Denne J, Enas G. Subgroup identification based on differential effect search—a recursive partitioning method for establishing response to treatment in patient subpopulations. Stat Med 2011;30:2601–21.
[10] Wasserstein RL, Lazar NA. The ASA statement on p-values: context, process, and purpose. Am Stat 2016;70:129–31.