Reproducibility of statistically significant phase III oncology trials: An In Silico meta-epidemiological analysis

Sherry, A.D.; Msaouel, P.; Miller, A.M.; Lin, T.A.; Jaoude, J.A.; Kouzy, R.; ... ; Ludmir, E.B.

European Journal of Cancer 226 (2025) 115596

Contents lists available at ScienceDirect

European Journal of Cancer

journal homepage: www.ejcancer.com

Original research

Reproducibility of statistically significant phase III oncology trials: An *In Silico* meta-epidemiological analysis

Alexander D. Sherry [a,b,*], Pavlos Msaouel [c,d], Avital M. Miller [a], Timothy A. Lin [a], Joseph Abi Jaoude [e], Ramez Kouzy [a], Adina H. Passy [a], Tomer Meirson [f], Nikolaos Ignatiadis [g], Zachary R. McCaw [h,i], Erik van Zwet [j], Ethan B. Ludmir [k,l]

[a] Department of Radiation Oncology, Division of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA
[b] Department of Radiation Oncology, Mayo Clinic, Rochester, MN, USA
[c] Department of Genitourinary Medical Oncology, Division of Cancer Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA
[d] Department of Translational Molecular Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA
[e] Department of Radiation Oncology, Stanford University, Stanford, CA, USA
[f] Davidoff Cancer Center, Rabin Medical Center-Beilinson Hospital, Petach Tikva, Israel
[g] Department of Statistics and Data Science Institute, University of Chicago, Chicago, IL, USA
[h] Insitro, South San Francisco, CA, USA
[i] Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[j] Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands
[k] Department of Gastrointestinal Radiation Oncology, Division of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA
[l] Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

ARTICLE INFO

Keywords:
Reproducibility
Replication
Cancer
Research
P values
Phase III
Clinical trials

ABSTRACT

Purpose: The conventional assumption that P values ≤ 0.05 imply reproducible effects has come under recent criticism. This concern is particularly relevant in oncology, as phase III oncology trials, which directly inform practice, are usually not repeated. Using advanced modeling techniques, we investigated the relationship between P values and reproducibility in oncology.

Methods: We obtained the signal-to-noise ratio distribution in phase III oncology using outcomes from 632 two-arm superiority trials enrolling 496,219 patients. With this distribution, we estimated successful replication probability as the probability that a replicate trial, having the same design, effect size, and standard error, would have a two-sided $P \leq 0.05$ and the same effect directionality as the original trial. We also estimated the following: the probability that the estimated effect had the same direction as the true effect (i.e., correct sign probability); the probability that the 95 % CI covered the true effect (i.e., coverage probability), and the ratio of the observed estimated effect to the true effect (i.e., exaggeration factor).

Results: The median exaggeration factor across all trials was 1.09 (IQR, 0.80–1.61). When $P \leq 0.05$ in the original trial, mean correct sign probabilities were $\geq$ 97 % and mean coverage probabilities were between 93 % and 96 %. However, effects at P of 0.05, 0.01, and 0.001 had mean replication probabilities of 43 % (95 % CI: 35–45 %), 60 % (95 % CI: 53–61 %), and 77 % (95 % CI: 71–79 %), respectively. For trials with an overall survival primary endpoint that led directly to regulatory approval, the median replication probability was 66 %. A user-friendly web interface is provided to facilitate estimation of replication probabilities of individual trials.

Conclusions: While the direction of observed effects is likely correct when $P \leq 0.05$, treatment effects at P of 0.05 in phase III oncology trials are unlikely to be replicated successfully. By itself, statistical significance should not be equated with high replication probability.

1. Introduction

The current interpretation of relative treatment effects in randomized trials focuses on the evaluation of P values, which estimate the probability of seeing data that is at least as extreme as the observed data when the null hypothesis and all other modeling assumptions are

* Correspondence to: Department of Radiation Oncology, Mayo Clinic, 200 1st St. SW, Rochester, MN, 55905, USA
E-mail address: sherry.alexander2@mayo.edu (A.D. Sherry).

https://doi.org/10.1016/j.ejca.2025.115596
Received 1 June 2025; Received in revised form 25 June 2025; Accepted 30 June 2025
Available online 4 July 2025

true [1,2]. In the context of randomized trials, and assuming that all other statistical modeling assumptions are correct, a small *P* value indicates that the observed differences in outcomes between the randomized arms are unlikely to have occurred by chance if the null hypothesis were true [3]. Thus, trials often use small *P* values as a means to interpreting superior treatment effects.

However, the use of *P* values to define relative treatment superiority has come under increasing criticism amid growing concerns that research across scientific disciplines has been increasingly difficult to reproduce [4–7]. Dichotomized statistical significance thresholds, most commonly a *P* value ≤ 0.05, have been challenged as both arbitrary and oversimplified, as *P* is a continuous statistic [8–10]. In particular, binary thresholds can blur the nuanced distinction between statistical significance vs clinical significance, whose interpretation requires additional information such as effect size, as has been recognized by multiple societies such as ASCO and ESMO [11,12]. A small *P* value does not directly confirm the presence of bona fide treatment differences or truth, and a small *P* value does not necessarily imply that the results would replicate in repeated trials [13]. Lastly, the relationship between the reproducibility and *P* values, especially for *P* values at or near 0.05, remains poorly understood [14,15]. This is especially true for late-phase oncology trials, whose results are often used to define the standard of care; however, such results are rarely validated in subsequent trials owing to the cost, resources, time investment, and a perceived lack of equipoise after the results of the first "positive" trial have been released.

Despite this growing body of literature, most phase III oncology rely on small *P* values to justify relative treatment superiority interpretations. In this context, researchers are increasingly recognizing that clinical trial results may not effectively translate to routine oncology clinical practice [16,17]. To address these concerns, we conducted a large-scale meta-epidemiologic analysis of modern randomized oncology trials to examine the relationship between *P* values and reproducibility. Using modern statistical techniques, we modeled the probability of successful replicability, defined as the probability of obtaining a *P* value ≤ 0.05 in a replicate study with the same direction of effects as the original study, assuming the same effect size, standard error, and trial design [18,19]. We created a user-friendly online interface to facilitate reproducibility estimation for trials of interest. We also estimate correct sign probability, coverage probability, and the exaggeration factor for phase III oncology trials.

2. Methods

This study did not require institutional review board approval, as all data were publicly available. We performed a meta-epidemiological analysis consistent with the modified PRISMA guidelines [20]. Randomized phase III oncology trials were identified using the ClinicalTrials.gov registry. ClinicalTrials.gov was searched in May 2024 using the following criteria: terms "cancer," phase "Phase 3", study results "With Results," and status "excluded: Not yet recruiting." There was no restriction on the date of trial publication. Trials were required to have a two-arm, superiority design evaluating a time-to-event primary endpoint (PEP). Trials with more than two arms or trials assessing other PEPs, such as quality of life, were excluded. Because effect size was necessary to estimate replication probability, trials that did not report the estimated effect of treatment on the PEP with a hazard ratio (HR) and 95 % CI obtained by a Cox regression were excluded (Fig. S1).

Trial-level characteristics, including the summary statistics for the PEP outcome, were manually curated into a standardized database (Supplemental methods). Successful replication probability was defined as the probability that a trial replication with exactly the same design, true effect size, sample size (both number of patients and number of events), and standard error, would obtain a two-sided *P* less than 0.05 with the estimated effect in the same direction as the original study. The same definition of successful replication probability was applied to all trials, whether positive or negative. Here, true effect size refers not to the estimated effect size observed in the trial, but the unobservable true relative treatment effect of the comparison in question. We obtained this information with the following approach. The z score, which is directly computable from summary statistics of the Cox regression, is the ratio of the estimated treatment effect to the standard error. The z score is the sum of the signal-to-noise (SNR) ratio, defined as the ratio of the true treatment effect to the standard error, plus a standard normal error term. Thus, after obtaining the distribution of z scores of the underlying trials, we applied the deconvolution trick, detailed in multiple previous publications, to obtain the distribution of the SNR [18,19,21–25]. Notably, the distribution of the SNR can be obtained even though the true treatment effect cannot be observed in any individual trial. In addition to the prior methods publications, details are provided in the **Supplement** (Fig. S2 and Table S1). With this distribution, we computed replication probability using previously defined methods (code included in the **Supplement**) [18,19]. We also estimated successful replication probability after multi-fold increases in sample size of the replication study (corresponding directly to multi-fold increases in event number). 95 % CIs for the replication probabilities were computed using the Dvoretzky–Kiefer–Wolfowitz F-localization approach (code included in the **Supplement**) [19,26].

We also estimated several other measures of interest using this distribution. The correct sign probability was defined as the probability that the estimated effect had the same direction as the true effect. The coverage probability was defined as the probability that the 95 % CI covered the true effect. The exaggeration factor was calculated as the ratio of the estimated effect observed in the study to the true effect. Analysis was completed using R v4.3.2 (R Core team, Vienna, Austria), Julia v1.11.1 (Cambridge, MA), and SAS v9.4 (Cary, NC).

3. Results

Of the 1176 trials we screened, we identified 632 two-arm superiority phase III trials enrolling a total of 496,219 patients (Fig. S1**).** Trials were published between 2002 and 2024 with a median publication year of 2017 (IQR, 2014 to 2020). The trials' characteristics and outcomes are shown in Table 1. Most trials were sponsored by industry (n = 547, 87 %). Median enrollment was 563 patients (IQR, 361–838). Surrogate PEPs (n = 376, 59 %) were more common than overall survival (OS) PEPs (n = 256, 41 %). Half the trials (n = 319, 50 %) were interpreted as positive, according to the pre-specified statistical criteria for the PEP. The coverage probabilities, exaggeration factors, successful replication probabilities, and correct sign probabilities for different *P* value strata are shown in Table 2.

Coverage probabilities (i.e., the probability of the 95 % CI containing the true effect) were high for most *P* value strata, and exceeded 95 % for trials with $P > 0.05$. For trials with *P* values less than 0.001, the mean coverage probability decreased to 93 %. Thus, although smaller *P* values, with narrower 95 % CI, reflect more certain estimates and have a much higher probability of replication, there is a small tradeoff, resulting in a slightly lower probability of including the true effect in the interval.

The median exaggeration factor was 1.09 (interquartile range [IQR], 0.80–1.61), suggesting that the effects of most of the phase III trials in our analysis were modestly overestimated. *P* values between 0.01 and 0.5 were associated with the highest median exaggeration factors. Notably, *P* values between 0.05 and 0.1, which are commonly interpreted as "marginal significance" or "trending to significance," were associated with a median exaggeration of 15 %, which was 107 % at the upper quartile. This means that for 25 % of phase III oncology trials that yielded *P* values between 0.05 and 0.1 (7 % of all trials), the estimated effect size may be more than twice larger than the true effect size. Exaggeration was similarly problematic for the *P* value stratum between 0.01 and 0.05, with median exaggeration of 13 % and an exaggeration of 69 % at the upper quartile.

The median successful replication probability across all trials was

Table 1
Trial-level characteristics and outcomes along with replication probability.

Characteristic	n (%)	Replication probability, median (IQR)
Disease site		
Hematologic	114 (18)	63 (26–93)
Breast	103 (16)	45 (20–74)
Gastrointestinal	107 (17)	43 (18–68)
Genitourinary	89 (14)	42 (19–79)
Thoracic	111 (18)	40 (14–84)
Other[a]	108 (17)	38 (15–82)
Stage		
Metastatic	396 (63)	46 (19–82)
Non-metastatic	114 (18)	28 (11–53)
Hematologic	114 (18)	63 (26–93)
Cooperative group-sponsored		
Yes	91 (14)	20 (10–45)
No	541 (86)	50 (20–85)
Industry-funded		
Yes	547 (87)	51 (20–85)
No	85 (13)	20 (11–42)
Primary endpoint		
Overall survival	256 (41)	33 (14–60)
Surrogate survival	376 (59)	58 (21–93)
Enrollment		
$\leq$ 200 patients	42 (7)	37 (16–82)
$>$ 200 patients	590 (93)	45 (18–81)
Publication year		
Prior to 2017	298 (47)	45 (20–83)
2017 or later	334 (53)	45 (17–80)
Primary outcome		
Positive	319 (50)	81 (60–96)
Negative	313 (50)	18 (11–29)
Findings led to FDA approval		
Yes	232 (37)	82 (59–97)
No	379 (60)	23 (12–46)
Therapy previously approved	21 (3)	50 (23–73)

Abbreviations: Ref (reference); FDA (Food and Drug Administration); IQR (interquartile range).
Successful replication probability was defined as the probability that a trial replication with exactly the same design, true effect size, sample size (both number of patients and number of events), and standard error, would obtain a two-sided P less than 0.05 with the estimated effect in the same direction as the original study.
[a] Other disease sites included central nervous system, endocrine, gynecologic, germ cell, head and neck, sarcoma, and skin histologies.

45 % (IQR, 18–81 %). Overall, this is unsurprising, since nearly half trials had observed *P* values $>$ 0.05 (corresponding to z scores less than 1.96), with median *P* of 0.04 (Fig. 1A-B; Fig. S3). For effects at *P* values of 0.05, the mean replication probability was only 43 % (95 % CI: 35–45 %) (Fig. 1C). For effects at *P* of 0.01, the mean replication probability was 60 % (95 % CI: 53–61 %) (Fig. 1C). The low replication probabilities for *P* values in this range likely reflect inherent variability of the *P* value, in addition to the exaggeration, of effects at this level. The mean replication probabilities did not exceed 75 % for *P* values $>$ 0.001 (Fig. 1C). Conversely, *P* values $\leq$ 0.001 showed mean replication probability of 92 % but were obtained in only 29 % of the included studies (Table 2).

For studies with strong evidence of effects (i.e., $P = 0.001$), a 2-fold increase in sample size of the replicate study relative to the original study increased the mean replication probability from 77 % (95 % CI: 71–79 %) to 91 % (95 % CI: 88–92 %) (Fig. 2). However, for studies with weak evidence (i.e., $P = 0.05$), 10-fold increases in sample size to the replication study relative to the original study were required to raise the mean replication probability to 85 %, though even this replicate sample size was associated with considerable uncertainty (95 % CI: 73–88 %). This finding suggests that treatment effect estimates at $P = 0.05$ are rarely reproducible unless enrollment numbers are inflated well beyond practicality (median enrollment of the original studies was 563 patients).

The replication probabilities of trials grouped by different characteristics are given in Table 1. Industry-sponsored trials had a higher median replication probability (51 %, IQR, 20–85 %) than non-industry-sponsored trials (20 %; IQR, 11–42 %). The median replication probability for validation trials of therapies already approved by the FDA was 50 % (IQR, 23–73 %). Few trials with OS as the PEP had high reproducibility, with the median replication probability being 33 % (IQR, 14–60 %). Trials with an OS PEP had lower median replication probabilities than trials with surrogate PEPs (58 %; IQR, 21–93 %), possibly due to lower number of events reported than in trials with surrogate PEPs and a lower treatment difference. However, even for trials with a positive OS PEP, the median replication probability was only 66 % (IQR, 53–83 %) (Fig. 3). Similarly, the median replication probability for trials with an OS PEP that directly led to Food and Drug Administration (FDA) approval was 66 % (IQR, 49–83 %) (Fig. 3). Given the interpretative value added by replication probability, we created an easy to-use standalone web page for calculating successful replication probabilities for current or future phase III trials (Fig. 4). This webpage leverages the signal-to-noise ratio distribution computed in this study for this estimation, and requires only the input of the summary statistics (HR and 95 % CI) for any given phase III oncology trial of interest.

As shown in Table 2, the correct sign probability (i.e., the probability that the treatment effect direction was correctly identified) was low for trials with *P* values greater than 0.10. For example, for *P* values between 0.10 and 0.50, the mean correct sign probability was 81 %. Large *P* values, especially those above 0.10, should therefore not be interpreted

Table 2
Mean coverage probability, median and interquartile range (IQR) of the exaggeration factor, mean replication probability, and mean correct sign probability for different *P* value strata.

***P*-value stratum**	**Proportion**	**Coverage probability, mean**	**Exaggeration factor, median (IQR)**	**Replication probability, mean**	**Correct sign probability, mean**
0.9–1	0.04	97 %	0.13 (0.05–0.64)	11 %	52 %
0.8–0.9	0.04	97 %	0.38 (0.19–1.96)	11 %	56 %
0.7–0.8	0.04	97 %	0.63 (0.31–2.99)	12 %	60 %
0.6–0.7	0.04	97 %	0.84 (0.42–3.79)	13 %	64 %
0.5–0.6	0.04	97 %	1.02 (0.52–4.08)	14 %	68 %
0.1–0.5	0.24	96 %	1.2 (0.71–3.28)	22 %	81 %
0.05–0.1	0.07	96 %	1.15 (0.81–2.07)	37 %	93 %
0.01–0.05	0.11	93 %	1.13 (0.85–1.69)	50 %	97 %
0.005–0.01	0.04	94 %	1.12 (0.88–1.53)	62 %	99 %
0.001–0.005	0.07	95 %	1.11 (0.9–1.46)	70 %	100 %
0–0.001	0.29	93 %	1.1 (0.96–1.29)	92 %	100 %

Successful replication probability was defined as the probability that a trial replication with exactly the same design, true effect size, sample size (both number of patients and number of events), and standard error, would obtain a two-sided P less than 0.05 with the estimated effect in the same direction as the original study. The correct sign probability was defined as the probability that the estimated effect had the same direction as the true effect. The coverage probability was defined as the probability that the 95 % CI covered the true effect. The exaggeration factor was calculated as the ratio of the estimated effect observed in the study to the true effect; an exaggeration factor $>$ 1 suggests that the estimated observed effect reported in the trial is exaggerated relative to the true effect.

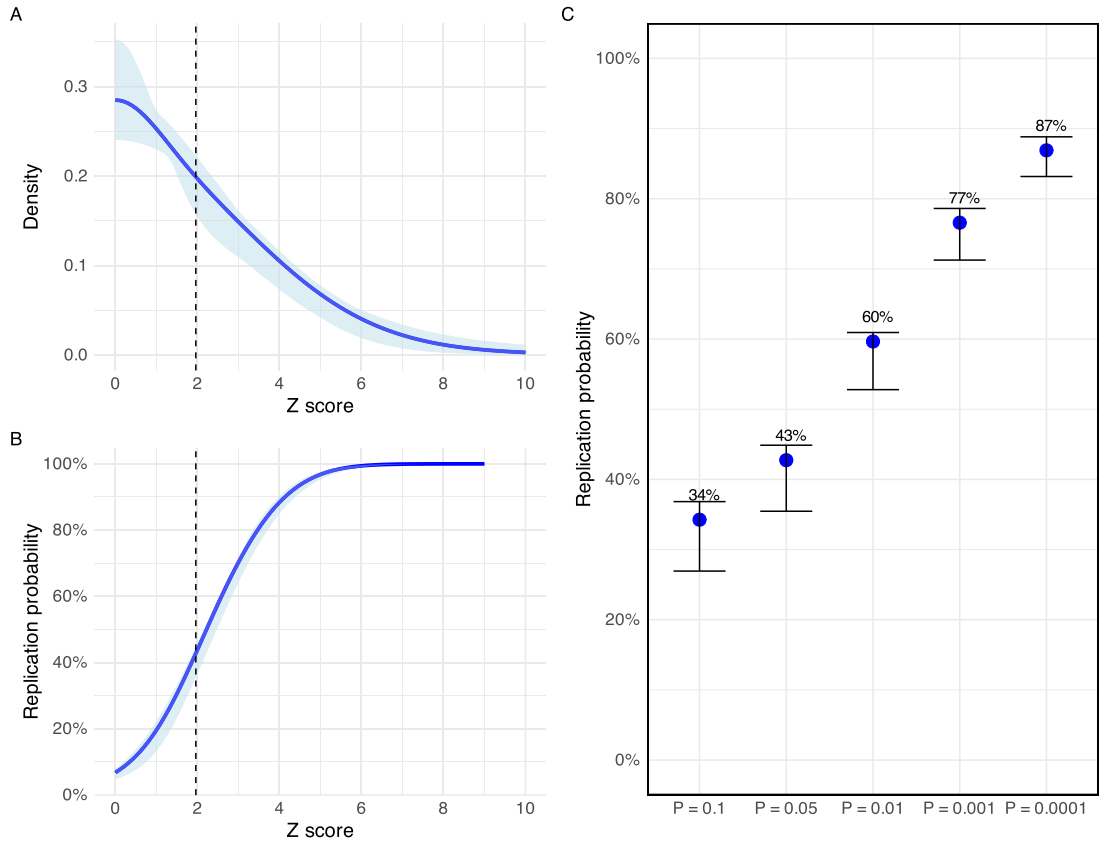

Fig. 1. (A) The distribution of the absolute z scores of the trials included in the present study. The blue line indicates the marginal density, and the shaded region indicates the 95 % CI bands. The dashed line shows z = 1.96, corresponding to $P = 0.05$. (B) Replication probabilities vs absolute z scores. The shaded region indicates the 95 % CI. The dashed line shows z = 1.96, corresponding to $P = 0.05$. (C) Mean replication probability estimated for the indicated P values. The bars represent the 95 % CIs.

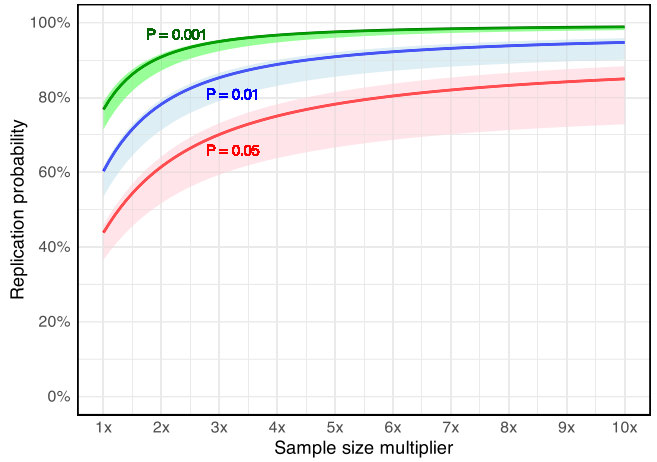

Fig. 2. Effects on replication probability of fold increases in sample size of the replication study relative to the original study, grouped according to initial P value. Lines indicate the mean point estimates, and the shaded areas indicate the 95 % CIs.

as evidence of a "signal" or "trend" in favor of treatment effects in a certain direction, consistent with prior work that large P values provide little information [27]. However, the mean correct sign probability was high for trials with small P values: 93 % for P between 0.05 and 0.1, 97 % for P between 0.01 and 0.05, and 99 % or greater for $P < 0.01$. Thus, sign errors (also known as type III errors or type S errors) appear quite unlikely when statistically significant results at $P < 0.05$ are obtained in phase III oncology trials [28].

4. Discussion

In this large-scale *in silico* study of modern phase III oncology trials, treatment effects at P of 0.05 were associated with low probabilities of successful replication. While the direction of observed effects is likely correct when $P \leq 0.05$, conventional statistical significance thresholds appear to be inadequate for reliably anticipating replication of treatment effect superiority in phase III oncology trials. Our assessment of the reproducibility of statistically significant phase III oncology trials offers timely and evidence-based insights for regulatory stakeholders, oncologists, and patients, aiding in the interpretation of current trials and informing the design of future phase III trials.

To our knowledge, this is the first study to estimate the reproducibility of phase III oncology trials. General randomized trials in medicine from the Cochrane database have been studied previously using our approach [18]. Compared to general randomized trials in medicine, phase III oncology trials are larger, more resource-intensive, subject to heightened scrutiny, and more likely to change practice. Therefore, as might be expected, our results suggest that phase III oncology trials are less prone to treatment effect exaggeration while demonstrating similar average coverage probability, higher average replication probability, and similar average correct sign probability for P value strata between 0.01 and 0.1 when compared to general randomized trials in medicine [18].

Our study adds to the growing body of literature carefully evaluating the information provided by P values. Other notable examples include the concept of an S value, which rescales P values directly to bits of information and reflects the degree of surprise one would expect from a

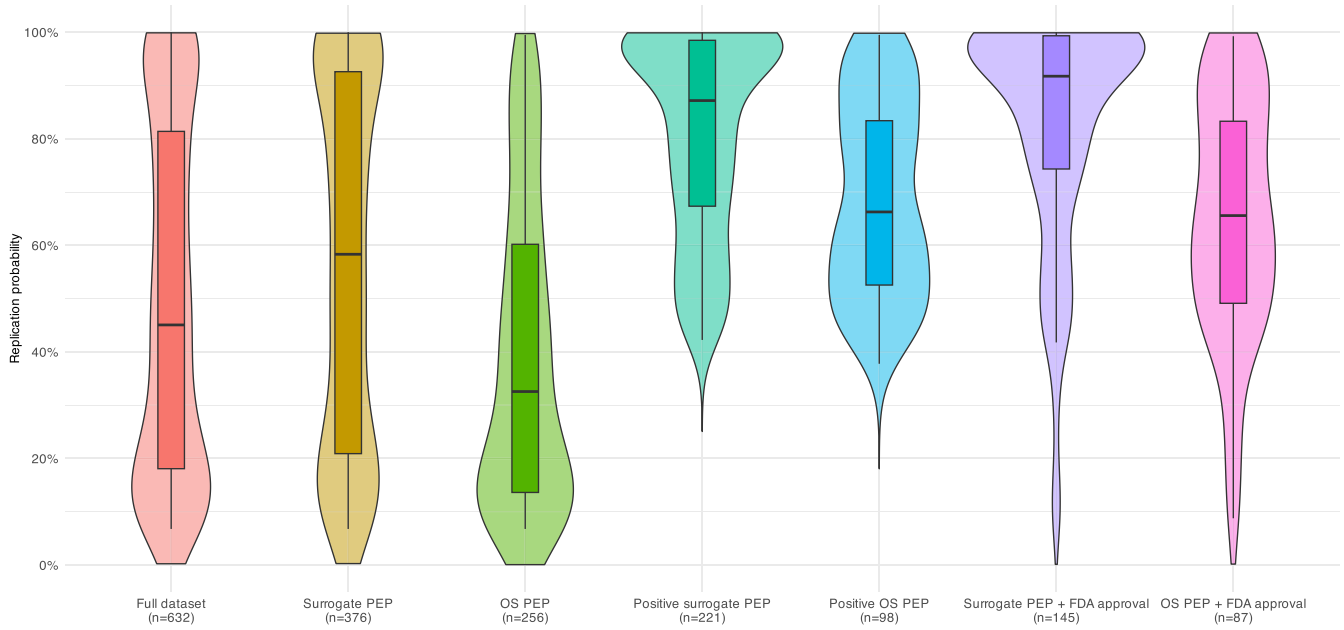

Fig. 3. Replication probabilities for all trials in the dataset and subgroups of interest. Boxplots show medians and interquartile ranges, and violin plots show the distribution of the data.

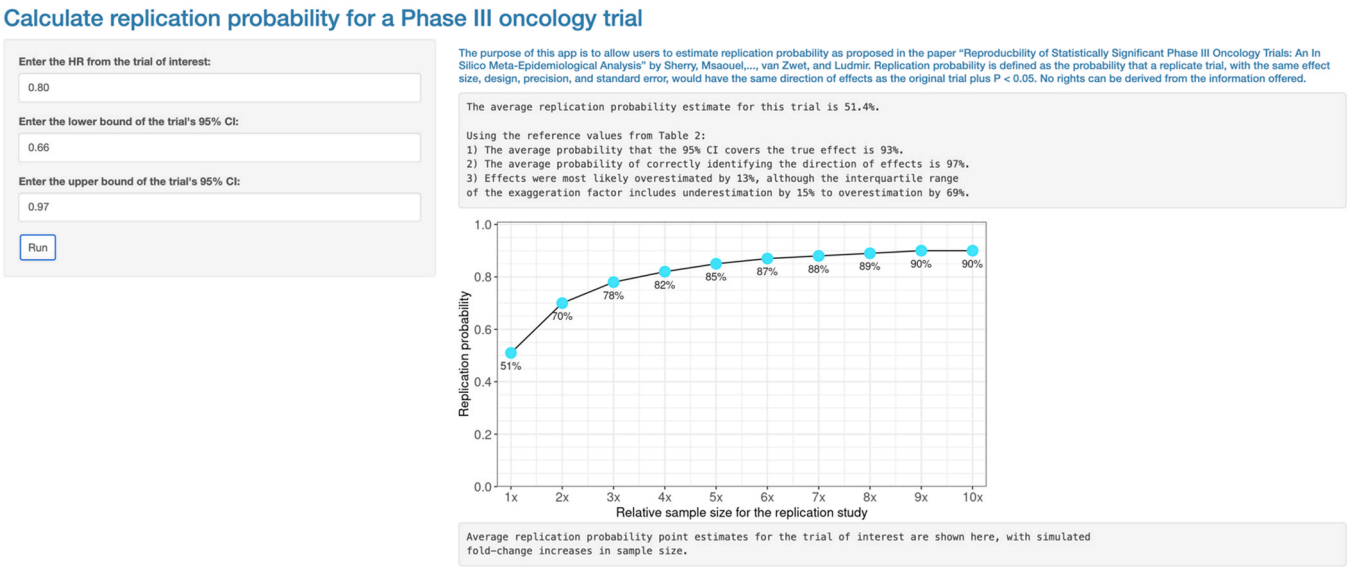

Fig. 4. Example of the interface of the application we created with Shiny to calculate replication probabilities for individual trials. The application is at https://alexandersherry.shinyapps.io/shinyapp_replication_probability/.

series of S tosses of a presumably fair coin [3,27,29]. A P value of 0.05 corresponds to $S = 4$ bits of information, suggesting that the findings of a trial resulting in a P value of 0.05 are only as surprising as 4 consecutive fair coin tosses all resulting in tails. Furthermore, a P value of 0.001 corresponds to 10 bits, which is more than twice the amount of information as the 4 bits of $P = 0.05$, consistent with our observations of improved reproducibility we observed for a P value of 0.001 compared with a P value of 0.05.

An initial implication from this work suggests that future studies, desiring improved replication probability, should consider reducing α beyond the conventional 0.05 error assertion rate when estimating sample size requirements and pre-specifying significance thresholds. However, there are numerous challenges for reducing α, such as increased trial costs, limited patient samples eligible for additional enrollments, and added time to trial completion. Moreover, reducing α even 10-fold to 0.005 would still only be associated with replication probabilities < 70 % [30]. Thus, although it is reassuring that effects at $P < 0.05$ are most likely in the correct direction, rather than reducing α to improve replication probability, this study more strongly argues that effects at low P values most commonly seen in phase III oncology trials is not a guarantee for effects likely to replicate. Rather than looking solely to P, other measures require greater attention for the evaluation of superiority, such as treatment effect size [31,32]. In oncology, stakeholders from the American Society of Clinical Oncology proposed a set of criteria for "minimum clinically important differences" [11]. In the context of replication probabilities, evaluation of treatment effect sizes remains imperative, because clinically trivial differences can still yield low P values and high replication probabilities, as demonstratable by the online webpage provided by our study [33,34]. Importantly, P values do not describe effect size, and low P values should not be interpreted as

representing clinically significant effects. To enhance *P* value interpretation, the probability of clinically relevant treatment effects may be directly estimated by Bayesian models [35,36]. Multiple online calculators are available to calculate such probabilities with summary statistics, if patient-level data are not available [24,37,38]. Moreover, the endpoint itself also matters considerably for understanding clinical relevance. In our study, replication probability was typically much greater for trials evaluating surrogate endpoints compared with trials evaluating OS. Surrogate PEPs, such as progression-free survival, may be less relevant to patients than OS or quality of life, and are frequently not well correlated with either of these patient-centric endpoints [39–42]. In some cases, the association between surrogate PEPs and the endpoints they are intended to represent may be questionable or even spurious, and trials with initially positive surrogate PEPs often subsequently show no differences in OS [43–46]. Our finding that trials with OS PEPs had significantly lower replication probabilities than trials with surrogate PEPs underscores the need to increase the power and precision of OS analyses and account for biases affecting OS estimation, as OS may have lower event rates in many trials compared with surrogate endpoints [8, 34,47,48].

The major limitation of this study is that modeling cannot substitute for actual trial replication, and we encourage trial validation through subsequent trials or post-marketing evaluations whenever possible. However, we recognize the considerable challenges inherent to replicating trials in oncology, where perceived equipoise is often lost after an initial positive study, funding is scarce, and current publication biases favor novelty over validation. Thus, our findings help to bridge the knowledge gap regarding clinical oncology research replication, and actively facilitate the interpretation of current and future phase III randomized trials. Moreover, our definition of replication probability, which assumes the exact same treatment effect, sample size, and standard error, is likely overoptimistic and generous; replication probability from conducting validation trials would most likely be lower than those estimated by our modeling definition [49]. We further caution that replication probabilities must be interpreted in the context of other factors. Such factors include but are not limited to effect size, absolute measures of relative treatment effects, endpoint selection, trial biases, covariates in the regression model, and the underlying assumptions of regression model [50,51]. Examples of trial biases, which may inflate replication probabilities, include substandard control arms, informative censoring, unblinded disease evaluation, and unaccounted post-progression therapies [47,52,53]. Additionally, therapies that result in severe toxicities must be weighed against the relevant benefits on a per-patient basis [34]. Accordingly, we discourage overly simplistic use of a "threshold" replication probability for determining whether treatment effects from a clinical trial are meaningful. Lastly, note that our replication probabilities are dependent only on the observed two-sided *P* values or, equivalently, absolute z-scores; thus, the correlations observed between replication probabilities and trial-level features are the same as those between two-sided *P* values (or z-scores) and trial-level features. Influences such as publication bias, if trials with lower *P* values were more likely to be published and thus represented in the dataset, may add bias to replication estimates, as well as other factors such as whether multivariable analysis was used for trial regression, whether informative censoring affected effect estimates, or whether analysis was conducted in the protocol-specified manner [51,54]. Furthermore, the accuracy of replication estimates may be influenced by scenarios where the assumptions underlying the effect estimation by Cox regression were not met in the original trial, such as a lack of proportional hazards [50].

Overall, phase III oncology trial with statistically significant findings at *P* of 0.05 reliably indicate the correct direction of treatment effects, but are unlikely to be successfully reproduced if an identical trial was repeated. While common in other fields, replication of phase III trials prior to regulatory approval is rare in oncology and unlikely to change for the foreseeable future. Although increasing the effective sample size can increase trial reproducibility, enrolling more patients may not always be feasible [55,56]. Meta-analyses of trials addressing a similar question are one approach to increasing confidence in the conclusion, although have their own limitations such as individual patient-level data availability. At the trial-level, efforts to improve reproducibility should include reports of long-term follow-up and careful multivariable regression modeling [57]. Adjustments for established prognostic factors in primary endpoint analyses should be used whenever possible, as they can increase information content, improving the power and efficiency of randomized trials [51]. Other potential strategies for increasing a trial's effective sample size include minimizing patient heterogeneity, employing biomarker enrichment to select patients most likely to benefit for the trial therapy, and enrolling patients at the highest risk for progression, albeit at the cost of a slower accrual rate. Consensus building through the engagement of stakeholders at multiple levels and reforming the regulatory approval process are essential steps in addressing the reproducibility crisis in late-phase oncology research.

Funding

This work was supported in part by the National Institutes of Health/National Cancer Institute through MD Anderson's Cancer Center Support Grant P30CA016672. Pavlos Msaouel and Ethan Ludmir are recipients of the Andrew Sabin Family Foundation Fellowship.

CRediT authorship contribution statement

Pavlos Msaouel: Writing – review & editing, Supervision, Resources, Methodology, Investigation, Conceptualization. **Ethan B. Ludmir:** Writing – review & editing, Supervision, Software, Resources, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Alexander D. Sherry:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Timothy A. Lin:** Writing – review & editing, Investigation, Data curation. **Avital M. Miller:** Writing – review & editing, Investigation, Data curation. **Nikolaos Ignatiadis:** Writing – review & editing, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Erik van Zwet:** Writing – review & editing, Visualization, Supervision, Software, Resources, Methodology, Investigation, Formal analysis, Conceptualization. **Zachary R. McCaw:** Writing – review & editing, Methodology, Investigation, Conceptualization. **Ramez Kouzy:** Writing – review & editing, Investigation, Data curation. **Joseph Abi Jaoude:** Writing – review & editing, Investigation, Data curation. **Tomer Meirson:** Writing – review & editing, Methodology, Investigation. **Adina H. Passy:** Writing – review & editing, Investigation, Data curation.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: **Alexander Sherry** reports honoraria from Sermo. Pavlos Msaouel reports honoraria for scientific advisory board membership for Mirati Therapeutics, Bristol-Myers Squibb, and Exelixis; consulting fees from Axiom Healthcare; non-branded educational programs supported by DAVA Oncology, Exelixis, and Pfizer; leadership or fiduciary roles as a Medical Steering Committee Member for the Kidney Cancer Association and a Kidney Cancer Scientific Advisory Board Member for KCCure; and research funding from Regeneron Pharmaceuticals, Takeda, Bristol-Myers Squibb, Mirati Therapeutics, and Gateway for Cancer Research (all unrelated to this manuscript's content). **Tomer Meirson** reports consulting fees from Purple Biotech. **Zachary McCaw** reports employment at Insitro (unrelated to this manuscript's content). **No other authors** report any conflicts of interest.

Acknowledgments

We thank Joe Munch in MD Anderson's Research Medical Library for editing the manuscript.

Data avail-ability

The code and data for this study are included in the Supplementary Materials. The authors created a standalone web page for calculating the replication probabilities for individual phase III oncology trials of interest: https://alexandersherry.shinyapps.io/shinyapp_replication_probability/.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ejca.2025.115596.

References

[1] Concato J, Hartigan JA. P values: from suggestion to superstition. J Invest Med Oct 2016;64(7):1166–71. https://doi.org/10.1136/jim-2016-000206.
[2] Wasserstein RL, Lazar NA. The ASA statement on p-values: context, process, and purpose. Am Stat 2016;70(2):129–33. https://doi.org/10.1080/00031305.2016.1154108.
[3] Msaouel P, Lee J, Thall PF. Interpreting randomized controlled trials. Cancers (Basel) 2023;15(19):4674.
[4] Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature 2019;567(7748):305–7.
[5] Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "p < 0.05". Am Stat 2019;73(sup1):1–19. https://doi.org/10.1080/00031305.2019.1583913.
[6] Ioannidis JP. Why most discovered true associations are inflated. Epidemiology Sep 2008;19(5):640–8. https://doi.org/10.1097/EDE.0b013e31818131e7.
[7] Ioannidis JPA. Why most published research findings are false. PLOS Med 2005;2(8):e124. https://doi.org/10.1371/journal.pmed.0020124.
[8] Msaouel P, Lee J, Thall PF. Making patient-specific treatment decisions using prognostic variables and utilities of clinical outcomes. Cancers (Basel) Jun 1 2021;13(11):2741. https://doi.org/10.3390/cancers13112741.
[9] Gelman A, Loken E. The statistical crisis in science. Am Sci 2014;102(6):460. https://doi.org/10.1511/2014.111.460.
[10] Andrade C. The P value and statistical significance: misunderstandings, explanations, challenges, and alternatives. Indian J Psychol Med May-Jun 2019;41(3):210–5. https://doi.org/10.4103/ijpsym.Ijpsym_193_19.
[11] Ellis LM, Bernstein DS, Voest EE, et al. American Society of Clinical Oncology perspective: raising the bar for clinical trials by defining clinically meaningful outcomes. J Clin Oncol 2014;32(12):1277–80. https://doi.org/10.1200/jco.2013.53.8009.
[12] Cherny NI, Dafni U, Bogaerts J, et al. ESMO-Magnitude of Clinical Benefit Scale version 1.1. Ann Oncol 2017;28(10):2340–66. https://doi.org/10.1093/annonc/mdx310.
[13] Altman N, Krzywinski M. Interpreting P values. Nat Methods 2017;14(3):213–4. https://doi.org/10.1038/nmeth.4210.
[14] Amrhein V, Trafimow D, Greenland S. Inferential statistics as descriptive statistics: there is no replication crisis if we don't expect replication. 2019/03/29 Am Stat 2019;73(sup1):262–70. https://doi.org/10.1080/00031305.2018.1543137.
[15] Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 2016;31(4):337–50. https://doi.org/10.1007/s10654-016-0149-3.
[16] Walters C, Harter ZJ, Wayant C, et al. Do oncology researchers adhere to reproducible and transparent principles? A cross-sectional survey of published oncology literature. BMJ Open Dec 31 2019;9(12):e033962. https://doi.org/10.1136/bmjopen-2019-033962.
[17] Wang SV, Sreedhara SK, Schneeweiss S, et al. Reproducibility of real-world evidence studies using clinical practice data to inform regulatory and coverage decisions. 2022/08/31 Nat Commun 2022;13(1):5126. https://doi.org/10.1038/s41467-022-32310-3.
[18] van Zwet E, Gelman A, Greenland S, Imbens G, Schwab S, Goodman SN. A new look at P values for randomized clinical trials. EVIDoa2300003 NEJM Evid 2024;3(1). https://doi.org/10.1056/EVIDoa2300003.
[19] Yang Y, van Zwet E, Ignatiadis N, Nakagawa S. A large-scale in silico replication of ecological and evolutionary studies. Nat Ecol Evol 2024;8(12):2179–83. https://doi.org/10.1038/s41559-024-02530-5.
[20] Murad MH, Wang Z. Guidelines for reporting meta-epidemiological methodology research. Evid Based Med 2017;22(4):139. https://doi.org/10.1136/ebmed-2017-110713.
[21] van Zwet E, Schwab S, Senn S. The statistical properties of RCTs and a proposal for shrinkage. Stat Med 2021;40(27):6107–17. https://doi.org/10.1002/sim.9173.
[22] van Zwet E, Schwab S, Greenland S. Addressing exaggeration of effects from single RCTs. Significance 2021;18(6):16–21. https://doi.org/10.1111/1740-9713.01587.
[23] van Zwet E, Tian L, Tibshirani R. Evaluating a shrinkage estimator for the treatment effect in clinical trials. Stat Med 2023;43(5):855–68. https://doi.org/10.1002/sim.9992.
[24] Sherry AD, Msaouel P, Kupferman GS, et al. Evidenced-based prior for estimating the treatment effect of phase III randomized trials in oncology. JCO Precis Oncol 2024;(8):e2400363. https://doi.org/10.1200/po.24.00363.
[25] Zwet E, Gelman A. A proposal for informative default priors scaled by the standard error of estimates. Am Stat 2022;76(1):1–9. https://doi.org/10.1080/00031305.2021.1938225.
[26] Ignatiadis N, Wager S. Confidence intervals for nonparametric empirical Bayes analysis. J Am Stat Assoc 2022;117(539):1149–66.
[27] Rafi Z, Greenland S. Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise. BMC Med Res Methodol 2020;20(1):244. https://doi.org/10.1186/s12874-020-01105-9.
[28] Gelman A, Carlin J. Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. Perspect Psychol Sci Nov 2014;9(6):641–51. https://doi.org/10.1177/1745691614551642.
[29] Greenland S, Mansournia MA, Joffe M. To curb research misreporting, replace significance and confidence by compatibility: A Preventive Medicine Golden Jubilee article. Prev Med Nov 2022;164:107127. https://doi.org/10.1016/j.ypmed.2022.107127.
[30] Benjamin DJ, Berger JO, Johannesson M, et al. Redefine statistical significance. Nat Hum Behav 2018;2(1):6–10. https://doi.org/10.1038/s41562-017-0189-z.
[31] Mark DB, Lee KL, Harrell Jr FE. Understanding the role of P values and hypothesis tests in clinical research. JAMA Cardiol Dec 1 2016;1(9):1048–54. https://doi.org/10.1001/jamacardio.2016.3312.
[32] McCaw ZR, Tian L, Wei J, et al. Choosing clinically interpretable summary measures and robust analytic procedures for quantifying the treatment difference in comparative clinical studies. Stat Med Dec 10 2021;40(28):6235–42. https://doi.org/10.1002/sim.8971.
[33] Hahn AW, Dizman N, Msaouel P. Missing the trees for the forest: most subgroup analyses using forest plots at the ASCO annual meeting are inconclusive. 17588359221103199 Ther Adv Med Oncol 2022;14. https://doi.org/10.1177/17588359221103199.
[34] Msaouel P, Lee J, Karam JA, Thall PF. A causal framework for making individualized treatment decisions in oncology. Cancers (Basel) 2022;14(16):3923. https://doi.org/10.3390/cancers14163923.
[35] Fornacon-Wood I, Mistry H, Johnson-Hart C, Faivre-Finn C, O'Connor JPB, Price GJ. Understanding the differences between bayesian and frequentist statistics. Int J Radiat Oncol Biol Phys 2022;112(5):1076–82. https://doi.org/10.1016/j.ijrobp.2021.12.011.
[36] Siddique J. Bayesian (re)-analyses of clinical trial data. EVIDe2200297 NEJM Evid 2023;2(1). https://doi.org/10.1056/EVIDe2200297.
[37] Wijeysundera DN, Austin PC, Hux JE, Beattie WS, Laupacis A. Bayesian statistical inference enhances the interpretation of contemporary randomized controlled trials. J Clin Epidemiol 2009;62(1):13–21.e5. https://doi.org/10.1016/j.jclinepi.2008.07.006.
[38] Lane D., Andrew B. Bayesian re-analysis of clinical trials. 2021. Accessed May 21, 2024. ⟨https://benjamin-andrew.shinyapps.io/bayesian_trials/⟩.
[39] Allen CJ, Snyder RA, Horn DM, et al. Defining priorities in value-based cancer care: insights from the alliance for clinical trials in Oncology National Cooperative Group Survey. JCO Oncol Pract 2023;19(10):932–8. https://doi.org/10.1200/op.23.00159.
[40] Haslam A, Hey SP, Gill J, Prasad V. A systematic review of trial-level meta-analyses measuring the strength of association between surrogate end-points and overall survival in oncology. Eur J Cancer 2019;106:196–211. https://doi.org/10.1016/j.ejca.2018.11.012.
[41] Hwang TJ, Gyawali B. Association between progression-free survival and patients' quality of life in cancer clinical trials. Int J Cancer Apr 1 2019;144(7):1746–51. https://doi.org/10.1002/ijc.31957.
[42] Kovic B, Jin X, Kennedy SA, et al. Evaluating progression-free survival as a surrogate outcome for health-related quality of life in oncology: a systematic review and quantitative analysis. JAMA Intern Med Dec 1 2018;178(12):1586–96. https://doi.org/10.1001/jamainternmed.2018.4710.
[43] Belin L, Tan A, De Rycke Y, Dechartres A. Progression-free survival as a surrogate for overall survival in oncology trials: a methodological systematic review. Br J Cancer 2020;122(11):1707–14. https://doi.org/10.1038/s41416-020-0805-y.
[44] Burzykowski T, Buyse M, Piccart-Gebhart MJ, et al. Evaluation of tumor response, disease control, progression-free survival, and time to progression as potential surrogate end points in metastatic breast cancer. J Clin Oncol J Am Soc Clin Oncol 2008;26(12):1987–92. https://doi.org/10.1200/jco.2007.10.8407.
[45] Chen EY, Haslam A, Prasad V. FDA Acceptance of Surrogate End Points for Cancer Drug Approval: 1992-2019. JAMA Intern Med 2020;180(6):912–4. https://doi.org/10.1001/jamainternmed.2020.1097.
[46] Liu ITT, Kesselheim AS, Cliff ERS. Clinical benefit and regulatory outcomes of cancer drugs receiving accelerated approval. JAMA 2024;331(17):1471–9. https://doi.org/10.1001/jama.2024.2396.
[47] Sherry AD, Msaouel P, Lin T, et al. Postprogression therapy and confounding for the estimated treatment effect on overall survival in phase III oncology trials. BMJ Oncol 2024;3(1):e000322. https://doi.org/10.1136/bmjonc-2024-000322.
[48] Berg SA, La Rosa S, Zhang T, et al. Impact of postprogression therapies on overall survival: Recommendations from the 2023 kidney cancer association think tank meeting. Urol Oncol 2025;43:135–46. https://doi.org/10.1016/j.urolonc.2024.10.022.
[49] Turner RM, Jackson D, Wei Y, Thompson SG, Higgins JP. Predictive distributions for between-study heterogeneity and simple methods for their application in

Bayesian meta-analysis. Stat Med Mar 15 2015;34(6):984–98. https://doi.org/10.1002/sim.6381.
[50] Lin TA, McCaw ZR, Koong A, et al. Proportional hazards violations in phase iii cancer clinical trials: a potential source of trial misinterpretation. Clin Cancer Res Oct 15 2024;30(20):4791–9. https://doi.org/10.1158/1078-0432.Ccr-24-0566.
[51] Sherry AD, Passy AH, McCaw ZR, et al. Increasing power in phase III oncology trials with multivariable regression: an empirical assessment of 535 primary end point analyses. JCO Clin Cancer Inf 2024;(8):e2400102. https://doi.org/10.1200/cci.24.00102.
[52] Hilal T, Sonbol MB, Prasad V. Analysis of control arm quality in randomized clinical trials leading to anticancer drug approval by the us food and drug administration. JAMA Oncol Jun 1 2019;5(6):887–92. https://doi.org/10.1001/jamaoncol.2019.0167.
[53] Rosen K, Prasad V, Chen EY. Censored patients in Kaplan-Meier plots of cancer drugs: An empirical analysis of data sharing. Eur J Cancer Dec 2020;141:152–61. https://doi.org/10.1016/j.ejca.2020.09.031.
[54] Hsu EJ, Lin TA, Dabush DR, et al. Association of differential censoring with survival and suboptimal control arms among oncology clinical trials. J Natl Cancer Inst 2024;116(6):990–4.
[55] Shapiro DD, Msaouel P. Challenges and considerations in modern adjuvant therapy trials in renal cell carcinoma: a call to power. Eur Urol Oct 28 2024. https://doi.org/10.1016/j.eururo.2024.10.020.
[56] Shapiro DD, Msaouel P. Adjuvant therapy for renal cell carcinoma: Finding the signal in the noise. Urol Oncol Jul 12 2024. https://doi.org/10.1016/j.urolonc.2024.06.018.
[57] Senn S. Seven myths of randomisation in clinical trials. Stat Med 2013;32(9):1439–50. https://doi.org/10.1002/sim.5713.