Survival-Inferred Fragility of Statistical Significance in Phase III Oncology Trials

Alexander D. Sherry[1], MD, Yufei Liu[2], MD PhD, Pavlos Msaouel[3,4], MD PhD, Timothy A. Lin[1], MD, MBA, Alex Koong[1], Christine Lin[1], Joseph Abi Jaoude[5], MD, Roshal R. Patel[6], MD, Ramez Kouzy[1], MD, Molly B. El-Alam[1], MPH, Avital M. Miller[1], BA, Mohannad Owiwi[7], MD, Jonathan Ofer[8], MD, David Bomze[9], MD MPH, Zachary R. McCaw[10,11], PhD, Tomer Meirson[8], MD PhD, Ethan B. Ludmir[12,13], MD

1. Department of Radiation Oncology, Division of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

2. Department of Radiation Oncology, City of Hope National Medical Center, Duarte, CA, USA

3. Department of Genitourinary Medical Oncology, Division of Cancer Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

4. Department of Translational Molecular Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

5. Department of Radiation Oncology, Stanford University, Stanford, CA, USA

6. Department of Radiation Oncology, Memorial Sloan-Kettering Cancer Center, New York, NY, USA

7. Jerusalem Mental Health Center, Eitanim Psychiatric Hospital, Jerusalem, Israel

8. Davidoff Cancer Center, Rabin Medical Center-Beilinson Hospital, Petach Tikva, Israel

9. Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel

10. Insitro, South San Francisco, CA, USA

11. Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

12. Department of Gastrointestinal Radiation Oncology, Division of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

13. Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

Correspondence to: Dr. Alexander Sherry, The University of Texas MD Anderson Cancer Center, 1400 Pressler St., Unit 1422, Houston, TX 77030, USA

Email: adsherry@mdanderson.org

Data availability: Deidentified reconstructed data for individual patients are available at the online repository Figshare and accessible via: https://figshare.com/articles/dataset/Reconstructed_survival_data_from_Phase_3_oncology_trials/26103268. Code for this study is included in the supplement.

Acknowledgments: We thank Laura Russell, scientific editor in The University of Texas MD Anderson Cancer Center's Research Medical Library, for editing the manuscript.

Previous presentations: Portions of this study were previously presented in October 2024 at the American Society for Radiation Oncology Annual Meeting in Washington DC.

Support: 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.

Authors' disclosures of potential conflicts of interest: 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 as 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). Zachary McCaw reports employment at Insitro (unrelated to this manuscript's content). Tomer Meirson reports consulting fees from Purple Biotech. No other authors report any conflicts of interest.

Conception and design: Alexander Sherry, Yufei Liu, Pavlos Msaouel, Timothy Lin, Jonathan Ofer, David Bomze, Zachary McCaw, Tomer Meirson, Ethan Ludmir.

Collection and assembly of data: Alexander Sherry, Timothy Lin, Alex Koong, Christine Lin, Joseph Abi Jaoude, Roshal Patel, Ramez Kouzy, Molly El-Alam, Avital Miller, Ethan Ludmir. **Financial support:** Ethan Ludmir. **Software and code:** Alexander Sherry, Yufei Liu, Timothy Lin, Mohannad Owiwi, Jonathan Ofer, David Bomze, Tomer Meirson, Ethan Ludmir. **Data analysis and interpretation**: All authors. **Manuscript writing:** ADS wrote the first draft. All authors revised the paper for critical intellectual content. **Final approval of manuscript:** All authors. **Accountable for all aspects of work:** All authors

Key words: phase III trials; cancer; fragility; statistical significance; P values

ABSTRACT

Background: Statistical significance currently defines superiority in phase III oncology trials. However, this practice is increasingly questioned. Here, we estimated the fragility of phase III oncology trials.

Methods: Using Kaplan-Meier curves for the primary endpoints of 230 two-arm superiority phase III oncology trials, we reconstructed data for individual patients. We estimated the survival-inferred fragility index (SIFI) by iteratively flipping the best responder from the experimental arm to the control arm ($SIFI_B$) until the interpretation was changed according to the significance threshold of each trial. Severe fragility was defined by SIFI ≤1%.

Results: This study included 230 trials enrolling 184,752 patients. The median number of patients required to change trial interpretation was 8 (interquartile range, 4 to 19) or 1.4% (interquartile range, 0.7% to 3%) per $SIFI_B$. Estimations of SIFI by multiple methods were largely consistent. For trials with an overall survival primary endpoint, the median $SIFI_B$ was 1% (IQR, 0.5% to 1.9%). Severe fragility was found in 87 trials (38%). As a continuous statistic, the original *P* value—but not its binary significance interpretation—was associated with fragility and severe fragility. Trials with subsequent FDA approval had lower odds of severe fragility. Lastly, the underlying survival model had differential effects on SIFI estimation.

Conclusions: Even among phase III oncology trials, which directly inform patient care, changes in the outcomes of few patients are often sufficient to change statistical significance and trial interpretation. These findings imply that current definitions of statistical significance used in phase III oncology are inadequate to identify replicable findings.

INTRODUCTION

Superiority interpretations in phase III oncology trials are currently governed by statistical significance.[1] Statistical significance is defined by achieving a *P* value in the survival model less than or equal to a predefined threshold.[2,3] Although statistical significance is essentially universal in phase III oncology research, the robustness of statistical significance as the chief determinant of treatment superiority has been increasingly questioned.[4,5] Statistical significance assigns diametrically opposing interpretations to *P* of 0.049 and *P* of 0.051 (when significance is defined as $P < 0.05$, the usual convention), but *P* is a continuous statistic, which can vary stochastically from sample to sample. The information encoded by *P* of 0.049 and 0.051 is nearly identical.[6-10]

To quantify the robustness of statistical significance, the survival-inferred fragility index (SIFI) has been proposed.[11,12] The SIFI is defined as the minimum number of individual patient outcomes needed to change the interpretation of statistical significance. Because oncology trials investigate survival, which is a composite of both the timing and occurrence of the event (i.e., alive or deceased), reliable estimation of fragility in oncology research also requires a concurrent evaluation of both patient events and the timing of those events.[13,14] Thus, to understand fragility in oncology, data for individual patients is required, which has been a major limitation in a field where patient-level data are rarely shared.[15]

Previously, Bomze and colleagues reconstruction survival data for individual patients in 45 phase III trials testing immunotherapy strategies and found a median SIFI of only 5 patients.[16] Building on this important work, we aimed to estimate fragility among a superset of 230 phase III oncology trials. The purpose of the present study was to provide an updated and more comprehensive characterization of the fragility or robustness of statistical significance specific to phase III oncology trials; to investigate multiple methods of evaluating SIFI; to compare the effects of different survival statistical models on SIFI; and to determine whether alternative approaches to clinical trial interpretation are needed.

METHODS

We performed a meta-epidemiological analysis of phase III oncology trials identified from ClinicalTrials.gov in February 2020 with no date limitations. Only 2-arm superiority trials were included in this study. Primary endpoints (PEPs) that were not published, were not time-to-event, or lacked a Kaplan-Meir plot with a number-at-risk table were excluded. Screening with these criteria yielded 337 trials evaluable for the reconstruction of individual patients' survival data. Methods of manual survival-data reconstruction using WebPlotDigitizer (Austin, TX) have been reported previously.[17,18] The reconstruction quality was defined as the absolute value of the natural logarithm of $HR_{recon}/HR_{reported}$, where HR_{recon} was the point estimate of the hazard ratio (HR) for the reconstructed data determined by Cox regression, and $HR_{reported}$ was the point estimate of the HR reported in the trial publication. Reconstructions in which this value was greater than 0.1 were excluded. Trials with proportional hazards violations in the PEP were also excluded, because the Cox proportional hazards model, used for the SIFI estimation, is unreliable in this setting.[19] After application of these criteria, 230 trials were eligible for inclusion in the study and further analysis, as reported previously.[17] Trial-level features were recorded in a standardized database. Enrollment was defined as the number of patients in the PEP analysis. Surrogate endpoints were defined as disease-related endpoints attempting to represent overall survival, consistent with other publications.[20] United States' Food and Drug Administration (FDA) approvals were evaluated as reported previously.[21] Institutional review board approval was not needed because the data were publicly available. This report conforms to the modified Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) for meta-epidemiological studies.[22]

The fragility of each trial was estimated using R v.4.4.2 (Vienna, Austria) with previously published methods, and the code is included in the **Supplement**.[16] SIFI values were estimated by counting the number of patients that needed to be iteratively flipped between treatment arms to change the original statistical significance interpretation of the PEP Cox proportional hazards regression (i.e., from positive to negative or vice versa). Statistical significance was defined uniquely for each trial as the threshold set by the trial. Flipping of the best (longest) survivors from the experimental arm to the control arm to calculate SIFI_B was used for the main analysis. Sensitivity analyses were performed by flipping the worst (shortest) survivors from the control arm to the experimental arm (SIFI_W), and flipping of the median survivors from the control arm to the experimental arm (SIFI_M). SIFI counts were computed using R v4.4.2 (Vienna, Austria). After the absolute value of the SIFI count was calculated, the index was normalized as a percentage of the total number of participants evaluated in the PEP analysis. Based on prior work, we defined severe fragility as being present when the SIFI was less than or equivale to 1% of the study enrollment.[16]

To explore the effects of the survival model on SIFI, SIFI, as conventionally estimated by Cox proportional hazards regression, was compared to SIFI estimated by three alternative survival models: restricted mean survival time (RMST), MaxCombo2, and MaxCombo3. RMST models estimate the difference in mean survival between treatment arms by integrating the area under the survival curve until the truncation time τ, defined here as the earlier of the last observed event from either arm.[19] MaxCombo2 incorporates the Fleming–Harrington log-rank statistic plus a weighted log-rank statistic for late separation of curves, and MaxCombo3 provides an additional weighting to MaxCombo2 to account for diminishing treatment effects.[23]

Continuous variables were summarized by median and interquartile range (IQR). Correlations between the SIFI and the initial *P* value were estimated using the Spearman's rank correlation coefficient. The chi-square test was used to test the association between categorical variables. The associations between trial-level features and the SIFI were evaluated using the Mann-Whitney U test. Binary logistic regressions tested the association between trial-level features and severe fragility to estimate odds ratios (OR). Two-sided *P* values with 95% CIs were calculated using SAS v9.4 (Cary, NC), and significance was defined as $P < 0.05$. Plots were created using Prism v10 (La Jolla, CA).

RESULTS

The study included 230 trials, published from 2005 to 2020 and enrolling 184,752 patients in total (**Table S1**). Most trials evaluated surrogate PEPs (n=140, 61%). The PEP was positive (i.e., the experimental arm was interpreted as demonstrating superiority to the control arm) in 120 trials (52%), and led to FDA approving the experimental therapy tested in 82 trials (36%).

In the overall dataset, the median SIFI_B, SIFI_W, and SIFI_M counts were 8 patients (IQR, 4 to 19 patients), 11 patients (IQR, 6 to 20 patients), and 17 patients (IQR, 9 to 30 patients), respectively (**Figure 1A**). As percentages of the number of patients studied in the PEP analyses, the median SIFI_B, $\mathrm{SIFI}_{W,}$ and SIFI_M percentage were 1.4% (IQR, 0.7% to 3%), 1.9% (IQR, 0.9% to 3.8%), and 3% (IQR, 1.3% to 6.0%), respectively (**Figure 1B**). The SIFI values estimated by each method were well-correlated, especially for the SIFI_W and SIFI_M values (**Figure 2**). Severe fragility (i.e., SIFI percentage ≤ 1% of the study enrollment) was detected in 87 trials (38%) per the SIFI_B values, 68 trials (30%) per the SIFI_W values, and 49 trials (21%) per the SIFI_M values.

Trials with overall survival PEPs were significantly more fragile (i.e., had lower SIFI percentages) than trials with surrogate PEPs (median SIFI_B percentages: 1.0% vs 1.8%, respectively; $P < 0.0001$) (**Figure 3A**). Similarly, compared with trials with surrogate

PEPs, trials with overall survival PEPs were associated with greater odds of severe fragility, as defined by the SIFI$_B$ value (OR, 2.33; 95% CI, 1.35 to 4.06; P = 0.003). This difference may have been related to the fact that trials claiming superiority appeared modestly less fragile (i.e., had higher SIFI values) than trials that did not claim superiority (median SIFI$_B$ percentages: 1.9% vs 1.1%, respectively; P = 0.0002) because surrogate PEPs were more likely to result in superiority interpretations than overall survival PEPs (64% vs 30%, respectively; P < 0.0001 per the chi-square test) (**Figure 3B**). However, in the context of severe fragility, there was only a weak signal that initial statistical significance (i.e., claims of superiority) was associated with reduced odds of severe SIFI$_B$ fragility (OR, 0.62; 95% CI, 0.36 to 1.06; P = 0.08). This association became inconclusive after adjusting for the potential confounding effects of PEP type on trial outcome (adjusted OR, 0.78; 95% CI, 0.44 to 1.37; P = 0.38). Interestingly, despite these findings for the interpretation of P, the value of P itself appeared closely associated with severe SIFI$_B$ fragility ($-\log_{10}(P)$: OR, 0.73; 95% CI, 0.63 to 0.83; P < 0.0001) and after adjustment for PEP type (adjusted OR, 0.75; 95% CI, 0.64 to 0.86; P = 0.0001). This result was consistent when fragility was estimated for each SIFI approach (**Figure S1**).

On the other hand, the differences in fragility were more pronounced in the comparison of trials that did versus did not lead to FDA approval. The median SIFI$_B$ percentage of trials leading to FDA approval was 3.1%, and the median SIFI$_B$ percentage of trials not leading to FDA approval was vs 1.1% (P < 0.0001) (**Figure 3C**). Trials with FDA approval were associated with lower odds of severe SIFI$_B$ fragility than trials without subsequent FDA approval (OR, 0.47; 95% CI, 0.26 to 0.83; P = 0.01). This association persisted after adjustment for PEP type (adjusted OR, 0.53; 95% CI, 0.29 to 0.96; P = 0.04). There was also a significant relationship between FDA approval and severe fragility as estimated using SIFI$_W$ values (adjusted OR, 0.52; 95% CI, 0.27 to 0.96; P = 0.04), although the effect was weaker for severe fragility as estimated using SIFI$_M$ values (aOR, 0.59; 95% CI, 0.28 to 1.17; P = 0.14). Thus, although the SIFI percentages were low across the full dataset, including among trials claiming superiority, the FDA approval process did appear to select for findings that appeared, on the whole, notably less fragile (i.e., the median fragility of trails receiving approval was approximately one third that of trials not receiving approval). Other characteristics were not associated with fragility, such as whether or not the trial was an immunotherapy study (median SIFI$_B$ percentage, 1.47% [IQR, 0.60% to 4.0%] for immunotherapy trials vs 1.40% [IQR, 0.7% to 2.9%] for trials testing other approaches; P = 0.80).

The effects of different survival models on estimating the SIFI values are shown in **Figure 4**. The SIFI$_B$ values were highest (i.e., least fragile) when estimated by the RMST model (P < 0.0001 vs Cox), and lowest (i.e., most fragile) when estimated by the MaxCombo approaches (P < 0.0001 and P = 0.03 for MaxCombo2 and MaxCombo3, respectively, vs Cox) (**Figure 4A**). The inverse was demonstrated for the SIFI$_W$ values, where MaxCombo2 and MaxCombo3 were found to be least fragile, when compared with Cox, and RMST most fragile (**Figure 4B**).

DISCUSSION

In this study, we estimated the fragility of phase III oncology trial results by leveraging reconstruction techniques to assemble a uniquely comprehensive dataset of 184,752 individual patient outcomes from 230 trials. We found evidence of fragility as estimated by the SIFI approach in most trials; the median SIFI count was 8 patients, or a median percentage of the study enrollment of 1.4%. Severe fragility was identified in 87 (38%) trials. Taken together, our results provide novel and quantitative insights into the problems of using statistical significance to interpret the results of phase III oncology trials. Alternative approaches to trial interpretation that do not rely solely on statistical significance are urgently needed.

We found that severe fragility, while related to the initial *P* value as a continuous parameter, does not appear to be closely related to whether the initial *P* value was deemed significant or not. In other words, both “positive” and “negative” trials, as defined by statistical significance, are susceptible to fragility, and a small series of changes to individual outcomes may readily reverse the trial interpretation in either direction. This finding is not necessarily surprising from a conceptual standpoint; the information provided by *P* values of 0.049 and 0.051 is essentially identical.[24] However, this empirical observation is concerning because it implies that the extent of both type I errors and type II errors are currently underestimated in phase III oncology trials.[17] However, when *P* is viewed as a continuous statistic rather than a binary outcome, which is a more appropriate approach, *P* serves well as a marker of fragility.[3] While trials that obtained FDA approval did seem to exhibit less fragility on average, suggesting the importance of the downstream regulatory approval process, fragility still remained problematic even for this subset of trials, as the median SIFI value was 3%.

In our study, different approaches to estimating SIFI reveal the strengthens and weaknesses of the concept of fragility. We found strong concordance between SIFI estimated according to flipping the best survivor, the worst (shortest) survivor, and the median survivor. Moreover, fragility appears to be markedly influenced by the choice of the underlying survivor model. RMST analyses, which compare the area under the survival curve, appeared less fragile when the longest-term survivors were flipped between arms, consistent with expectations that late-occurring drops will hold less influence on area under the curve calculations. Conversely, MaxCombo models, which weight for late separation of the curves or late diminishing treatment effects, appeared more fragile when flipping the best survivors. We found the inverse to be the case when focusing SIFI on the earliest aspects of the survival curve in the SIFI_W models. Thus, the underlying data from an individual trial and the underlying survival model could considerably influence the results of fragility estimations, and each should be interpreted as offering distinct and unique information on the behavior of the underlying survival data.

Alternative approaches to clinical trial interpretation beyond statistical significance are urgently needed, and ultimately, the regulatory process, as perhaps implied by our data, appears most strongly poised to affect such change.[25] The European Society of Medical Oncology and the American Society of Clinical Oncology have proposed that evaluations of treatment effect superiority should include assessments of effect size.[26-28] Effect sizes describe the relative magnitude of benefit to patients and adds important clinical meaning to the interpretation of survival data.[29] Notably, because *P* values do not convey effect sizes (although they are often perceived in this way), treatment effects with very low *P* values (e.g., $P < 0.001$) may be associated with only marginal effect sizes.[7,30] Quantifying the probability of achieving clinically meaningful effect sizes in phase III oncology trials is a highly valuable strategy for trial interpretation, and would complement the use of statistical significance.[17,31] Lastly, it is essential to not only consider the challenges of trial interpretation, the generalizability of the trial, and caveats of trial design, but to also take into account the characteristics and needs of individual patients—such as patient values and patient-specific risks—when making clinical decisions.[32,33]

Caution is warranted when interpreting the present study for several reasons. This study's findings are subject to the limitations of reconstructed survival data.[18] Although trial-specific significance levels were used to mimic the conditions of the original studies, all reconstructed regressions were univariable. In contrast, most phase III oncology trials use multivariable regressions for the PEP because multivariable regressions have greater power and efficiency than univariable regressions.[34,35] Thus, the SIFI modeling approach may overestimate fragility for trials that were initially statistically significant, because lower *P* values, which reflect more information than higher *P* values, may be obtained when strongly prognostic covariates are included in the PEP regression model.[35] To reduce this bias from this risk, our analysis excluded studies where the HR based on reconstructed data differed from the original HR by more than 0.1 on the logarithmic scale. Like any single summary measure, the fragility index has limitations, and we do not suggest that it can provide a stand-alone alternative to the p-value for making decisions regarding treatment efficacy.[16] There is no consensus approach among researchers regarding the definitions of SIFI and severe fragility, and so we used multiple approaches to examine SIFI. Despite these limitations, we propose that the fragility index provides an intuitive metric for shedding light on the instability of clinical trial results.

Although studying SIFI is ultimately *in silico*, one of the key strengths of this study, compared to a pure-simulation study, is that actual patient outcomes were used to estimate fragility. Thus, our findings here have direct and immediate relevance to phase III oncology trials, which frequently change the standard of care. Building on a growing literature conveying the limitations of statistical significance criteria, this study provides new, quantitative, and easily understandable insights into the severe fragility of many late-phase trials. To improve the reliability of the evidence generated in oncology trials, alternative strategies for interpreting clinical trials beyond statistical significance are urgently needed.

FIGURE LEGENDS

Figure 1. The survival-inferred fragility index (SIFI) for the primary endpoints of phase III oncology trials according to SIFI_B, SIFI_W, and SIFI_M. (A) The absolute numbers of patients and (B) the percentages of the number of patients studied in the primary endpoint analyses are shown. Bars represent medians and interquartile ranges.

Figure 2. Different methods of calculating survival-inferred fragility index (SIFI) percentages provide highly concordant results. The Spearman's rank correlation coefficient is shown. In each figure, the solid line is the best fit univariable regression. The dashed line is the line of identity.

Figure 3. Trials claiming superiority, studying surrogate endpoints, and receiving FDA approval tend to have larger survival-inferred fragility indices (SIFI). The association of fragility, as measured by SIFI_B, with trial-level features: (A) Primary endpoint type; (B) statistical significance; (C) Subsequent FDA approval. *P* by Mann-Whitney U test. Bars represent medians and interquartile ranges.

Figure 4. The survival-inferred fragility index (SIFI) estimates are influenced by the underlying survival model. (A) SIFI_B and (B) SIFI_W estimated using Cox regression, restricted mean survival time (RMST), MaxCombo2 (MC2), and MaxCombo3 (MC3).

REFERENCES

1. Lin TA, Sherry AD, Ludmir EB. Challenges, Complexities, and Considerations in the Design and Interpretation of Late-Phase Oncology Trials. *Semin Radiat Oncol*. 2023;33(4):429-437.

2. Concato J, Hartigan JA. P values: from suggestion to superstition. *J Investig Med*. Oct 2016;64(7):1166-71. doi:10.1136/jim-2016-000206

3. Wasserstein RL, Lazar NA. The ASA Statement on p-Values: Context, Process, and Purpose. *The American Statistician*. 2016/04/02 2016;70(2):129-133. doi:10.1080/00031305.2016.1154108

4. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature*. 2019;567(7748):305-307.

5. Benjamin DJ, Berger JO, Johannesson M, et al. Redefine statistical significance. *Nature Human Behaviour*. 2018/01/01 2018;2(1):6-10. doi:10.1038/s41562-017-0189-z

6. Goodman SN. Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy. *Annals of Internal Medicine*. 1999/06/15 1999;130(12):995-1004. doi:10.7326/0003-4819-130-12-199906150-00008

7. 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. doi:10.1007/s10654-016-0149-3

8. Greenland S. Divergence versus decision P-values: A distinction worth making in theory and keeping in practice: Or, how divergence P-values measure evidence even when decision P-values do not. *Scandinavian Journal of Statistics*. 2023/03/01 2023;50(1):54-88. doi:https://doi.org/10.1111/sjos.12625

9. Greenland S. Invited Commentary: The Need for Cognitive Science in Methodology. *Am J Epidemiol*. Sep 15 2017;186(6):639-645. doi:10.1093/aje/kwx259

10. Boos DD, Stefanski LA. P-Value Precision and Reproducibility. *Am Stat*. 2011;65(4):213-221. doi:10.1198/tas.2011.10129

11. Walsh M, Srinathan SK, McAuley DF, et al. The statistical significance of randomized controlled trial results is frequently fragile: a case for a Fragility Index. *Journal of Clinical Epidemiology*. 2014;67(6):622-628. doi:10.1016/j.jclinepi.2013.10.019

12. Johnson KW, Rappaport E, Shameer K, Glicksberg BS, Dudley JT. fragilityindex: an R package for statistical fragility estimates in biomedicine. *bioRxiv*. 2019:562264.

13. Del Paggio JC, Tannock IF. The fragility of phase 3 trials supporting FDA-approved anticancer medicines: a retrospective analysis. *The Lancet Oncology*. 2019;20(8):1065-1069. doi:10.1016/S1470-2045(19)30338-9

14. Bomze D, Meirson T. A critique of the fragility index. *The Lancet Oncology*. 2019/10/01/ 2019;20(10):e551. doi:https://doi.org/10.1016/S1470-2045(19)30582-0

15. Naudet F, Sakarovitch C, Janiaud P, et al. Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: survey of studies published in The BMJ and PLOS Medicine. *BMJ*. 2018;360:k400. doi:10.1136/bmj.k400

16. Bomze D, Asher N, Hasan Ali O, et al. Survival-Inferred Fragility Index of Phase 3 Clinical Trials Evaluating Immune Checkpoint Inhibitors. *JAMA Netw Open*. Oct 1 2020;3(10):e2017675. doi:10.1001/jamanetworkopen.2020.17675

17. Sherry AD, Msaouel P, Kupferman G, et al. Towards Treatment Effect Interpretability: A Bayesian Re-analysis of 194,129 Patient Outcomes Across 230 Oncology Trials. *[preprint] medRxiv*. 2024:2024.07.23.24310891. doi:10.1101/2024.07.23.24310891

18. Guyot P, Ades AE, Ouwens MJ, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Med Res Methodol*. 2012;12:9. doi:10.1186/1471-2288-12-9

19. 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):4791-4799. doi:10.1158/1078-0432.Ccr-24-0566

20. Chen EY, Haslam A, Prasad V. FDA Acceptance of Surrogate End Points for Cancer Drug Approval: 1992-2019. *JAMA Intern Med*. 2020;180(6):912-914. doi:10.1001/jamainternmed.2020.1097

21. Abi Jaoude J, Kouzy R, Ghabach M, et al. Food and Drug Administration approvals in phase 3 Cancer clinical trials. *BMC Cancer*. 2021;21(1):695. doi:10.1186/s12885-021-08457-5

22. Murad MH, Wang Z. Guidelines for reporting meta-epidemiological methodology research. *Evidence Based Medicine*. 2017;22(4):139. doi:10.1136/ebmed-2017-110713

23. Mukhopadhyay P, Ye J, Anderson KM, et al. Log-Rank Test vs MaxCombo and Difference in Restricted Mean Survival Time Tests for Comparing Survival Under Nonproportional Hazards in Immuno-oncology Trials: A Systematic Review and Meta-analysis. *JAMA Oncol*. Sep 1 2022;8(9):1294-1300. doi:10.1001/jamaoncol.2022.2666

24. van Zwet E, Gelman A, Greenland S, Imbens G, Schwab S, Goodman SN. A New Look at P Values for Randomized Clinical Trials. *NEJM Evidence*. 2024;3(1):EVIDoa2300003. doi:doi:10.1056/EVIDoa2300003

25. Wasserstein RL, Schirm AL, Lazar NA. Moving to a World Beyond “p□<□0.05”. *The American Statistician*. 2019/03/29 2019;73(sup1):1-19. doi:10.1080/00031305.2019.1583913

26. Del Paggio JC, Sullivan R, Schrag D, et al. Delivery of meaningful cancer care: a retrospective cohort study assessing cost and benefit with the ASCO and ESMO frameworks. *The Lancet Oncology*. 2017/07/01/ 2017;18(7):887-894. doi:https://doi.org/10.1016/S1470-2045(17)30415-1

27. Cherny NI, Dafni U, Bogaerts J, et al. ESMO-Magnitude of Clinical Benefit Scale version 1.1. *Annals of Oncology*. 2017/10/01/ 2017;28(10):2340-2366. doi:https://doi.org/10.1093/annonc/mdx310

28. 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. doi:10.1200/jco.2013.53.8009

29. 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):6235-6242. doi:10.1002/sim.8971

30. Msaouel P, Lee J, Thall PF. Interpreting Randomized Controlled Trials. *Cancers (Basel)*. 2023;15(19):4674.

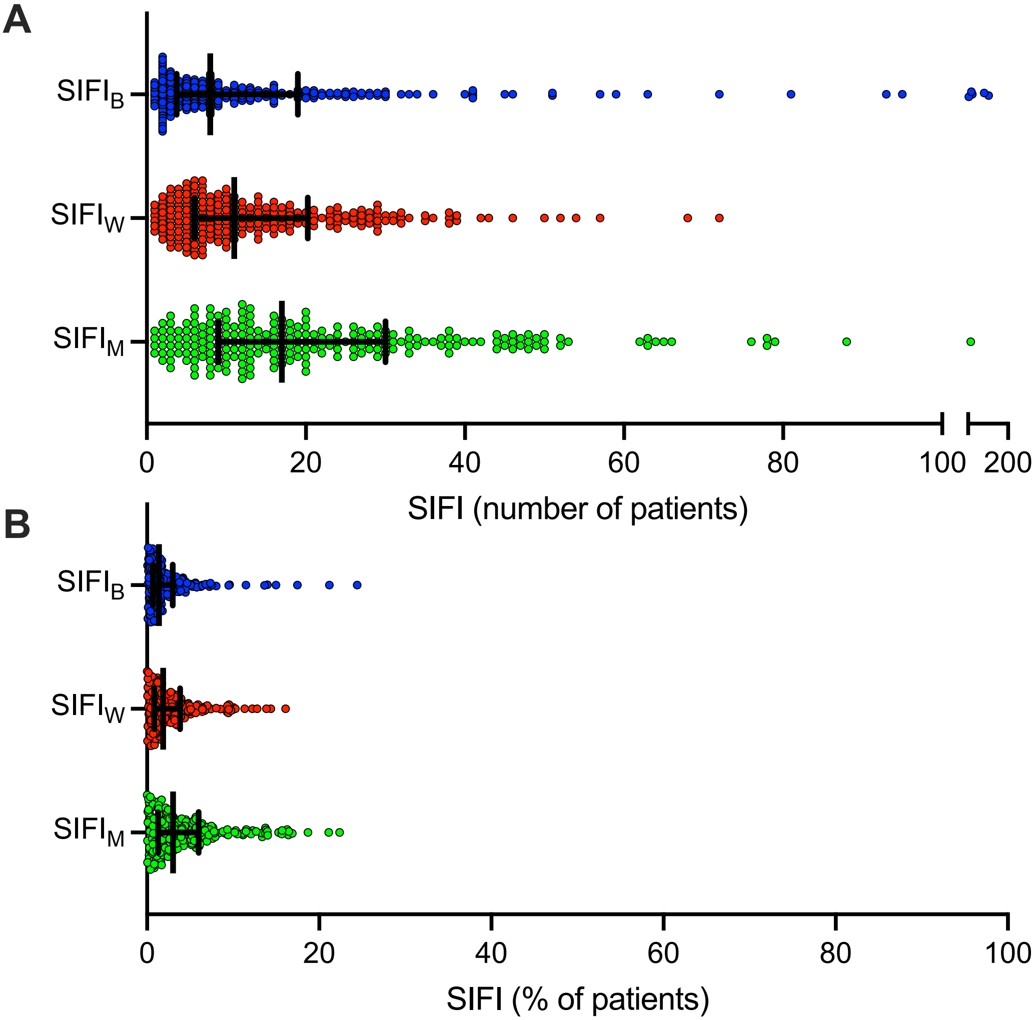
31. Sherry AD, Msaouel P, Kupferman GS, et al. Evidenced-Based Prior for Estimating the Treatment Effect of Phase III Randomized Trials in Oncology. *JCO Precis Oncol*. 2024;8:e2400363. doi:10.1200/po.24.00363

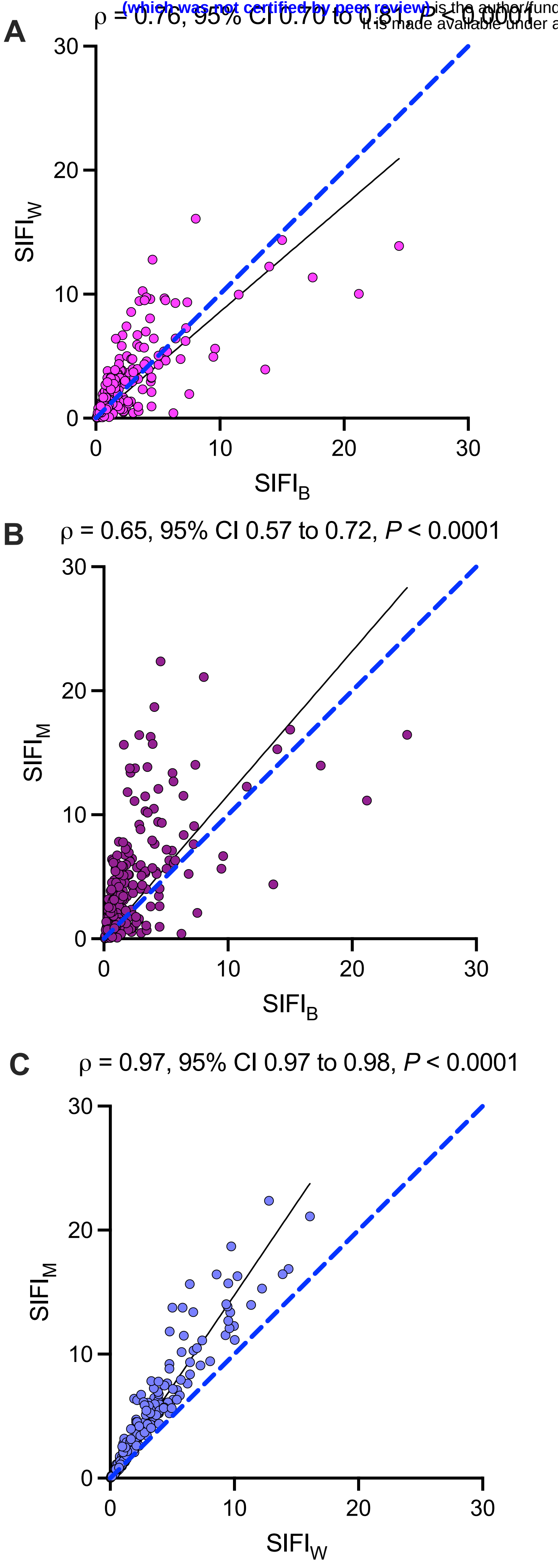
32. Msaouel P, Lee J, Karam JA, Thall PF. A Causal Framework for Making Individualized Treatment Decisions in Oncology. *Cancers (Basel)*. 2022;14(16):3923. doi:10.3390/cancers14163923

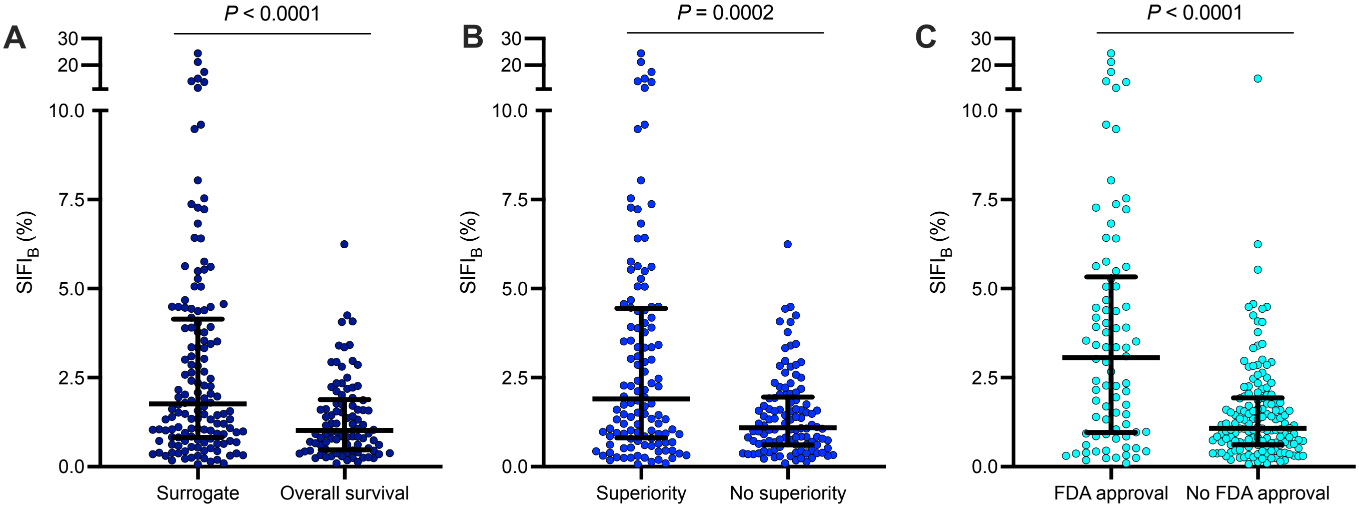
33. Msaouel P, Lee J, Thall PF. Making Patient-Specific Treatment Decisions Using Prognostic Variables and Utilities of Clinical Outcomes. *Cancers (Basel)*. Jun 1 2021;13(11):2741. doi:10.3390/cancers13112741

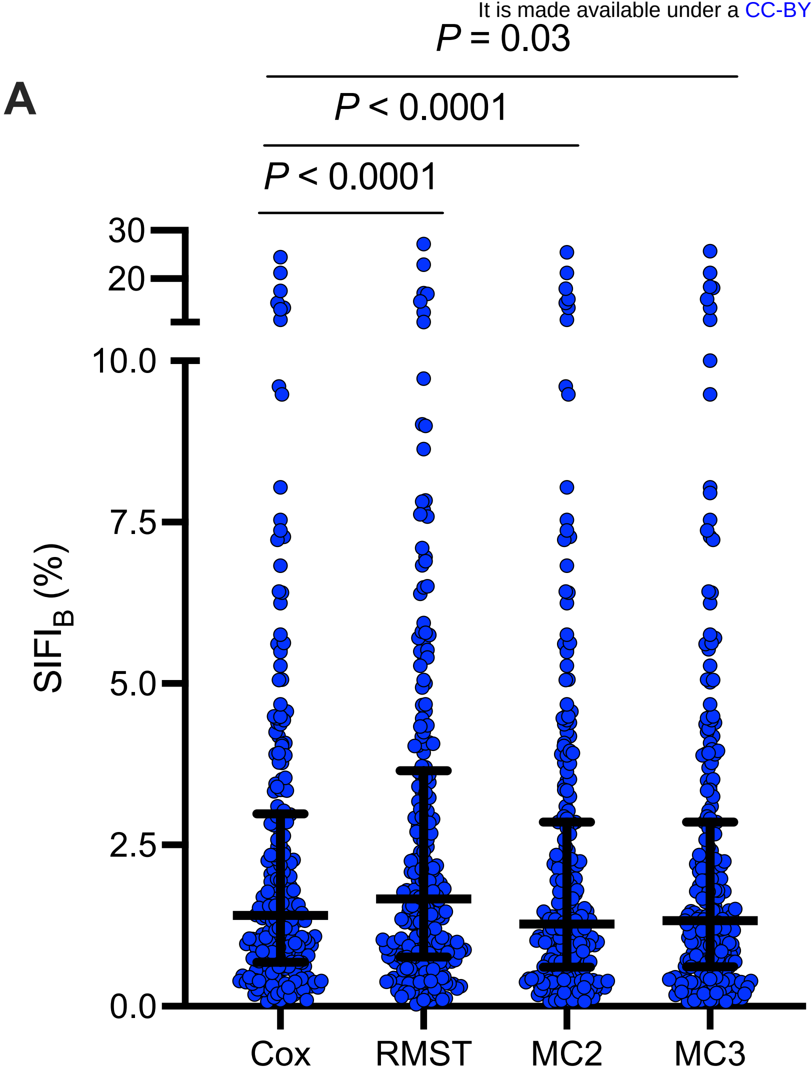
34. Sherry AD, Passy AH, McCaw ZR, et al. Increasing Power in Phase III Oncology Trials With Multivariable Regression: An Empirical Assessment of 535 Primary End Point Analyses. *JCO Clin Cancer Inform*. 2024;8:e2400102. doi:10.1200/cci.24.00102

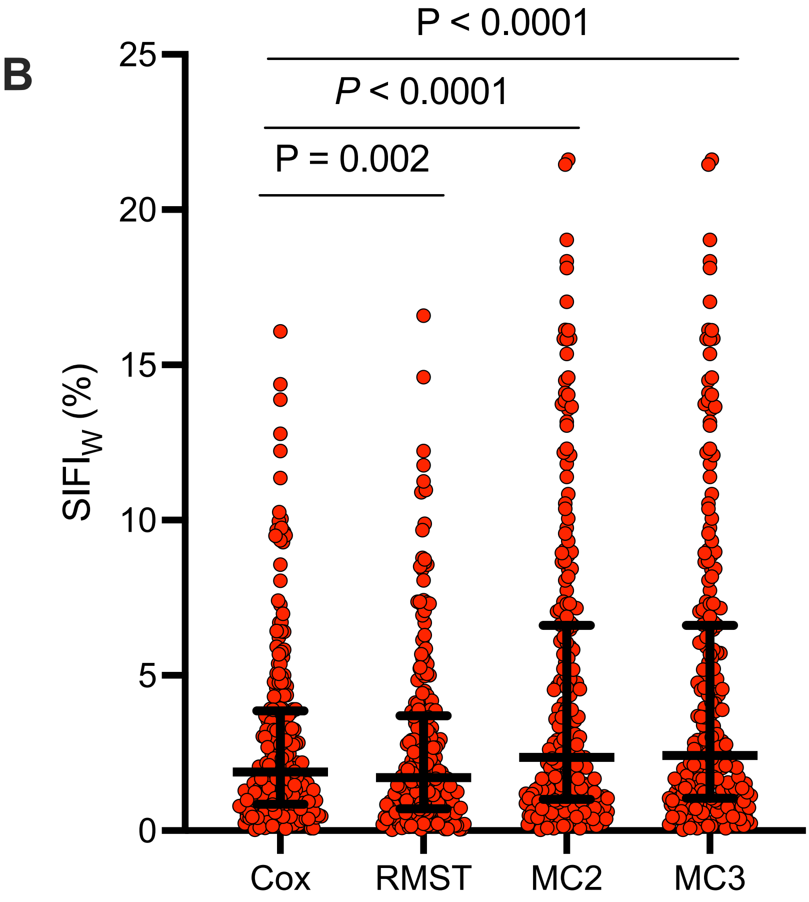
35. Senn S. Seven myths of randomisation in clinical trials. *Stat Med*. 2013;32(9):1439-50. doi:10.1002/sim.5713


A
SIFI$_B$
SIFI$_W$
SIFI$_M$
0
20
40
60
80
100
200
SIFI (number of patients)
B
SIFI$_B$
SIFI$_W$
SIFI$_M$
0
20
40
60
80
100
SIFI (% of patients)


A
ρ = 0.76, 95% CI 0.70 to 0.81, P < 0.0001
SIFI_W
SIFI_B
0
10
20
30
B
ρ = 0.65, 95% CI 0.57 to 0.72, P < 0.0001
SIFI_M
SIFI_B
0
10
20
30
C
ρ = 0.97, 95% CI 0.97 to 0.98, P < 0.0001
SIFI_M
SIFI_W
0
10
20
30


A
B
C
P < 0.0001
P = 0.0002
P < 0.0001
$SIFI_B$ (%)
30
20
10.0
7.5
5.0
2.5
0.0
Surrogate
Overall survival
Superiority
No superiority
FDA approval
No FDA approval


A
P = 0.03
P < 0.0001
P < 0.0001
30
20
10.0
7.5
5.0
2.5
0.0
$SIFI_B$ (%)
Cox
RMST
MC2
MC3


B
P < 0.0001
P < 0.0001
P = 0.002
25
20
15
10
5
0
$SIFI_W$ (%)
Cox
RMST
MC2
MC3