

Statistical Robustness and Follow-up Completeness of Survival Outcomes in Phase III Breast Cancer Trials

Yiru Hou

Cancer Hospital of Chinese Academy of Medical Sciences

Dongyan Liu

Cancer Hospital of Chinese Academy of Medical Sciences

Xuejing Zhang

Cancer Hospital of Chinese Academy of Medical Sciences

Siqi Wang

Cancer Hospital of Chinese Academy of Medical Sciences

Congtian Wu

Cancer Hospital of Chinese Academy of Medical Sciences

Jiani Yuan

Henan Cancer Hospital

Xinxin Yan

Cancer Hospital of Chinese Academy of Medical Sciences

Lanwei Guo

Henan Cancer Hospital

Huiyao Huang

Cancer Hospital of Chinese Academy of Medical Sciences

Ning Li

`ncctrials@cicams.ac.cn`

Cancer Hospital of Chinese Academy of Medical Sciences

Research Article

Keywords: breast cancer, phase III randomized controlled trials, statistical fragility, survival-inferred fragility index, meta-epidemiologic study

Posted Date: September 2nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10544025/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Breast cancer is the most common malignancy worldwide. Survival outcomes from phase III RCTs guide approvals, guidelines, and treatment strategies, making statistical stability important. HRs, CIs, and P values show statistical significance but do not directly quantify robustness to outcome changes in a few patients.

Methods

We conducted a systematic review-based meta-epidemiologic study of phase III breast cancer drug RCTs (Jan 1, 2015-Dec 31, 2025) from PubMed, Embase, and Cochrane Library. Eligible endpoints were prespecified primary survival and key secondary time-to-event outcomes with reconstructable Kaplan–Meier curves and number-at-risk data. Primary outcome: SIFI; secondary: SFQ and $WCLFU \geq |SIFI|$, with subgroup analyses.

Results

We included 157 endpoint outcomes: 83 (52.9%) positive and 74 (47.1%) negative. Median SIFI was 14 (IQR, 5 to 34.5) for positive and -17 (-45 to -5.2) for negative. Median SFQ was 0.0157 (IQR 0.0093–0.0353) and -0.0157 (-0.0219 to -0.0058), respectively. Among 117 with WCLFU data, 69 (59.0%) had $WCLFU \geq |SIFI|$; blinding was independently associated with higher odds. Associated factors for SIFI/SFQ differed by positivity, including sample size, disease setting, therapeutic line, endpoint, and HR category.

Conclusions

Survival conclusions showed substantial statistical instability; in over half of evaluable outcomes, WCLFU met or exceeded the fragility boundary. Statistical significance alone may overstate stability. Reporting SIFI, SFQ, and follow-up completeness alongside HRs, 95% CIs, and P values may support more transparent interpretation for treatment decisions and regulatory evaluation.

Introduction

Breast cancer is one of the most common malignancies among women worldwide, and treatment evidence in this field is rapidly evolving.[1–4] Results from phase III randomized trials not only inform drug approvals and guideline recommendations, but also contribute to health technology assessment and reimbursement decisions, ultimately shaping routine clinical practice.[2, 5–7] This is particularly relevant amid continuing advances in CDK4/6 inhibitors, HER2-targeted therapies, AKT pathway inhibitors, immunotherapy, and antibody-drug conjugates.[2, 8–11] Although clinicians and regulators

commonly rely on hazard ratios, confidence intervals, and P values to assess trial results, these measures do not indicate how easily a conclusion could be overturned by outcome changes in a small number of patients.[12–14]

Fragility analysis provides a complementary perspective for evaluating the robustness of statistical findings from randomized controlled trials.[14–16] The fragility index and fragility quotient were initially applied to binary outcomes, and subsequent methodological work extended this logic to time-to-event analyses, leading to the development of the survival-inferred fragility index (SIFI) and survival fragility quotient (SFQ).[15–18] Previous fragility studies have shown that some positive findings from phase III oncology RCTs that have influenced clinical practice may be less robust than expected, with changes in the outcomes of only a small number of patients potentially altering the conclusions.[18–20] Breast cancer trials warrant particular attention because commonly used survival endpoints, including OS, PFS, DFS, EFS, and IDFS, can be affected by the number of events and censoring, which may ultimately influence drug approvals and treatment selection.[5, 21]

Previous oncology studies have raised similar concerns. Trials supporting FDA-approved anticancer drugs have been reported to have low fragility indices,[19] and SIFI analyses of immune checkpoint inhibitor trials and broader phase III oncology trials further showed that statistically significant survival findings may be fragile in some settings.[17, 18] However, breast cancer-specific evidence remains limited. We conducted a systematic review-based meta-epidemiologic analysis of published phase III randomized drug trials in breast cancer to evaluate the distribution and associated factors of SIFI and SFQ, and to examine whether withdrawals of consent or losses to follow-up reached or exceeded the fragility boundary.

Methods

Study design

This meta-epidemiologic analysis was conducted using published reports of phase III randomized breast cancer trials. Reporting followed the PRISMA 2020 statement [22], and the study protocol was registered in PROSPERO (CRD420251182325). Because this study used only published RCT reports and reconstructed time-to-event data, ethics approval was not required.

Data sources and study selection

PubMed, Embase, and the Cochrane Library were systematically searched for studies published from January 1, 2015, to December 31, 2025. Eligible studies were full-text original reports of phase III randomized drug trials in breast cancer. Eligible outcomes included the prespecified primary survival endpoint and key secondary time-to-event efficacy endpoints when Kaplan–Meier curves and number-at-risk tables were available in the main text or supplementary materials to support reconstruction of time-to-event data. Detailed inclusion and exclusion criteria and the full search strategy are provided in the

Supplementary Appendix. Two investigators independently screened titles, abstracts, and full texts according to prespecified criteria. Disagreements were resolved by discussion or adjudication by a third investigator. The study selection process is shown in Fig. 1.

Data extraction and reconstruction

We used a standardized data extraction form to systematically collect trial design characteristics, clinical features, and outcome data from each trial report. If a trial reported multiple eligible survival endpoints, each endpoint was treated as an independent analytical unit for assessing the robustness of results. However, each trial was counted only once when summarizing trial-level characteristics.

For studies that provided Kaplan–Meier survival curves, we reconstructed approximate individual patient-level time-to-event data from the published survival plots. Curve coordinates were first extracted using Digitizelt version 2.6.6, and patients' survival times and censoring status were then reconstructed in R version 4.5.3 using established methods proposed by Guyot et al. and Wei and Royston. [23, 24] When multiple eligible endpoints were available from the same trial, reconstruction and analysis were performed separately for each endpoint.

Fragility outcomes and key definitions

The prespecified primary outcome was SIFI. For statistically positive outcomes, SIFI was defined as the minimum number of patients whose outcomes would need to be reassigned to change a statistically significant result to non-significance.[17, 18] For statistically negative outcomes, negative SIFI represented the minimum number of patients whose outcomes would need to be reassigned for a non-significant result to cross the threshold for statistical significance.[18] The secondary outcome was SFQ, calculated as SIFI divided by the total sample size. SFQ was used to account for differences in trial size and to enable standardized comparisons of fragility across trials of different scales.[15, 18] Smaller absolute values of SIFI and SFQ indicated greater statistical fragility.

WCLFU was defined as the total number of participants who discontinued the study because of withdrawal of consent or loss to follow-up. To assess whether follow-up incompleteness exceeded the tolerance implied by statistical fragility, we constructed a binary outcome defined as $WCLFU \geq |SIFI|$. [14, 19]

Statistical analysis

All analyses were performed using R version 4.5.3. We prespecified four outcome-reassignment strategies to estimate SIFI: flip_best, flip_worst, clone_best, and clone_worst. The flip_best method was selected as the primary analytical approach because it estimates the smallest number of patient-level reassignments required to alter the statistical significance interpretation of a trial result.[17, 18]

Continuous variables were summarized as medians with interquartile ranges, and categorical variables as frequencies with percentages. Comparisons of SIFI and SFQ between two groups were performed using the Mann-Whitney U test, whereas comparisons across three or more subgroups were performed using the Kruskal-Wallis test.

Univariable logistic regression was used to identify trial-level factors associated with $WCLFU \geq |SIFI|$. Variables were included in the multivariable logistic regression model based on clinical relevance or statistical associations suggested by univariable analyses. Effect estimates were reported as odds ratios (ORs) with 95% confidence intervals. All statistical tests were two-sided, and $P < 0.05$ was considered statistically significant.

Results

Study selection and outcome-level characteristics

A total of 59,553 database records were identified (PubMed, $n = 3,254$; Embase, $n = 25,773$; Cochrane Library, $n = 30,526$). After removal of 37,986 duplicates, 21,567 records remained; after initial cleaning, 4,330 records entered title/abstract screening. Subsequently, 291 breast cancer-related articles were assessed in full text. Ultimately, 145 articles were included in the analysis, corresponding to 133 trial projects.

The final analytical dataset included 157 trial endpoint outcomes, of which 98 were from industry-sponsored studies and 59 from investigator-initiated studies. Most outcomes were from superiority trials (145/157, 92.4%), multi-regional studies (127/157, 80.9%), and open-label or non-blinded reports (106/157, 67.5%). Outcomes from non-metastatic or curative-intent settings (81/157, 51.6%) and advanced or metastatic settings (76/157, 48.4%) were nearly balanced. PFS was the most common endpoint (64/157, 40.8%), followed by DFS (49/157, 31.2%), IDFS (22/157, 14.0%), OS (14/157, 8.9%), and other time-to-event endpoints (8/157, 5.1%). Complete outcome-level characteristics and fragility metrics are provided in **Supplementary Table 1**.

Overall distribution of SIFI and SFQ

Positive and negative outcomes were nearly balanced in number, but both groups showed substantial dispersion in statistical robustness. Among the 157 outcomes, 83 (52.9%) were statistically positive and 74 (47.1%) were statistically negative. The median SIFI was 14 (IQR, 5 to 34.5) for positive outcomes and -17 (IQR, -45 to -5.2) for negative outcomes. The median SFQ was 0.0157 (IQR, 0.0093 to 0.0353) for positive outcomes and -0.0157 (IQR, -0.0219 to -0.0058) for negative outcomes. The ranked distribution of outcome-level SIFI values among outcomes with $|SIFI| \leq 10$ is shown in **Fig. 2**.

Figure 2. Distribution of outcome-level SIFI among endpoints. Note. Only endpoints with $|SIFI| \leq 10$ are shown.

Subgroup analyses of fragility metrics

Among positive outcomes, absolute SIFI increased with sample size ($P < 0.001$), was greater in non-metastatic/curative-intent than advanced/metastatic settings (median SIFI, 41 vs 8; $P < 0.001$), differed across therapeutic line with the largest margins in perioperative studies (median SIFI, 43; $P < 0.001$), and varied strongly by endpoint, with IDFS showing the highest median SIFI (79) and OS the lowest (5.5; $P < 0.001$). Positive-result SFQ did not differ significantly by publication year or endpoint, but varied by hazard-ratio category ($P < 0.001$). Key subgroup contrasts are shown in **Fig. 3**. Complete subgroup results are provided in **Supplementary Table 2**.

Among negative outcomes, SIFI and SFQ were likewise heterogeneous. Negative outcomes in non-metastatic/curative-intent settings were farther from reversal than those in advanced/metastatic disease (median SIFI, -31 vs -4; $P < 0.001$), and both SIFI and SFQ varied significantly by sample size, disease setting, therapeutic line, endpoint, and hazard-ratio category. For SFQ, investigator-initiated outcomes were proportionally less fragile than industry-sponsored outcomes (median SFQ, -0.0201 vs -0.0129; $P = 0.004$). Negative DFS and IDFS outcomes were generally harder to reverse than negative PFS or OS outcomes. Key subgroup contrasts are shown in **Fig. 3**. Complete subgroup results are provided in **Supplementary Table 3**.

Figure 3. Key subgroup differences in absolute SIFI. Note. Boxplots show absolute SIFI by therapeutic line, endpoint type, and disease setting among positive outcomes (A–C) and negative outcomes (D–F).

WCLFU relative to the absolute SIFI threshold

WCLFU data were available for 117 outcomes. In 69 of 117 outcomes (59.0%), WCLFU was greater than or equal to the absolute fragility threshold, meaning that the number of participants not fully captured during follow-up was at least as large as the minimum patient-level perturbation needed to change the statistical significance interpretation. This finding does not invalidate the original trial results, but indicates that incomplete follow-up should be considered when interpreting the stability of HRs and P values. This pattern affected 30 of 54 positive outcomes and 39 of 63 negative outcomes with available WCLFU information, as shown in **Fig. 4**.

Figure 4. WCLFU relative to the absolute SIFI threshold. note: $|SIFI| \leq 50$ and $WCLFU \leq 50$, Outcomes outside the displayed range are not shown for readability.

In univariable analysis, investigator-initiated outcomes and single-region studies had lower odds of $WCLFU \geq |SIFI|$, whereas blinded trials had higher odds. In the multivariable model, blinding remained the only independently significant factor, with an adjusted OR of 3.84 (95% CI, 1.31–11.24; $P = 0.014$). Study type (adjusted OR, 0.48; 95% CI, 0.17–1.38; $P = 0.172$), study scope (adjusted OR, 0.53; 95% CI, 0.18–1.55; $P = 0.244$), and cancer setting (adjusted OR, 1.30; 95% CI, 0.50–3.37; $P = 0.594$) were no longer statistically significant after adjustment. Adjusted regression results are provided in **Supplementary Table 4**; sparse subgroup signals, especially TNBC, should be interpreted cautiously.

Discussion

This meta-epidemiologic analysis shows that the statistical interpretation of phase III breast cancer survival outcomes is not fully captured by whether a trial meets its endpoint, the magnitude of the hazard ratio, or the nominal P value.[9, 10] These conventional measures remain indispensable, but they do not indicate how many patient-level survival assignments would need to change before a statistically positive result becomes non-significant, or a negative result crosses the significance threshold.[17, 18] By applying SIFI and SFQ, we found substantial heterogeneity in robustness across contemporary phase III breast cancer trials. Notably, in a large proportion of outcomes with available follow-up data, WCLFU was greater than or equal to |SIFI|, indicating that incomplete follow-up was numerically comparable to, or greater than, the patient-level margin required to alter statistical significance interpretation. This observation should not be interpreted as evidence that the original trial conclusions were incorrect; rather, it highlights that some conclusions may rest on a narrower statistical buffer than is apparent from HRs and P values alone. Therefore, robustness metrics and follow-up completeness may provide useful complementary context for evaluating the evidential strength of phase III breast cancer trials.

Which trials are more fragile cannot be determined from SIFI alone. As SIFI is an absolute patient-count metric, it is strongly influenced by sample size, event accrual, endpoint maturity, and follow-up duration. SFQ therefore provides an important complementary perspective by normalizing SIFI to the total sample size, allowing robustness to be compared more fairly across trials of different scales. SIFI is useful for communicating the number of patient-level changes required to alter statistical interpretation, whereas SFQ indicates whether that number represents a small or large proportion of the randomized population. In our analysis, larger trials, non-metastatic or curative-intent settings, perioperative treatment, and DFS/IDFS endpoints generally showed higher absolute SIFI values, but these patterns should not be interpreted as evidence that such trials are uniformly more robust in a proportional sense.

When follow-up incompleteness was incorporated through the threshold $WCLFU \geq |SIFI|$, blinding was the only factor independently associated with this outcome after adjustment. This association does not imply that blinding directly increases fragility; it more likely reflects differences in trial design, endpoint ascertainment, reporting practices, follow-up procedures, or patient retention. These findings suggest that robustness assessment should not rely solely on trial size, HR magnitude, or nominal P value, and that follow-up quality control, patient retention, and transparent reporting of missing follow-up should receive particular attention in trials whose WCLFU approaches or exceeds the fragility margin.

Having shown that follow-up incompleteness may numerically exceed the fragility margin in a substantial proportion of outcomes, we further explored how statistical robustness relates to downstream clinical interpretation and guideline adoption. Guideline inclusion was defined as whether the corresponding regimen or experimental strategy was incorporated into at least one major breast cancer guideline by the prespecified cutoff date. Fragility should not be interpreted as synonymous with downstream clinical importance or guideline adoption. In an exploratory mapping of statistical significance, normalized net robustness, sample size, and guideline inclusion, outcomes incorporated

into major breast cancer guidelines generally clustered in regions with stronger conventional evidence signals, including larger sample sizes and more extreme P values. However, guideline inclusion was not restricted to outcomes with the largest fragility margins, and some guideline-included outcomes had WCLFU values that approached or exceeded the corresponding |SIFI| threshold. This exploratory visualization is shown in **Supplementary Fig. 1**. This pattern is clinically plausible because guideline recommendations are based on a broader evidentiary judgment, including effect size, endpoint relevance, disease context, biological plausibility, safety, unmet need, treatment availability, and consistency with the totality of evidence.[2–4] Therefore, SIFI, SFQ, and follow-up completeness should be viewed as complementary transparency metrics rather than determinants of guideline eligibility. Their main value is to identify practice-changing results for which the statistical significance rests on a narrower patient-level margin than is apparent from HRs and P values alone.

This framework may also provide context for statistically negative trials. A negative outcome with a small negative SIFI may lie close to the statistical significance boundary, suggesting that longer follow-up, more complete event ascertainment, or updated analyses may still be informative. However, such findings should not be interpreted as implying that additional follow-up would necessarily convert a negative trial into a positive one; rather, they indicate that the statistical interpretation remains sensitive to event accrual and data completeness.

From a clinical and regulatory perspective, these findings highlight the importance of interpreting follow-up completeness in relation to the fragility margin. For pivotal trials supporting approval, indication expansion, or guideline adoption, excessive WCLFU may reduce confidence in the stability of a survival conclusion when the number of incompletely followed participants is comparable to or exceeds |SIFI|. This does not mean that such trial conclusions are invalid, but it suggests that censoring patterns, reasons for withdrawal, between-arm balance in missing follow-up, and follow-up maturity should be examined more explicitly.[21] In clinical practice, fragility metrics may be particularly informative when treatment effects are modest, P values are close to the significance threshold, or practice-changing recommendations are based on survival endpoints with limited patient-level margins. Differences between perioperative and advanced-disease trials further suggest that SIFI and SFQ should be interpreted according to trial context: large curative-intent studies may appear more robust in absolute SIFI terms, whereas proportional robustness and follow-up completeness may tell a more nuanced story. The observed association between blinding and $WCLFU \geq |SIFI|$ should also be interpreted cautiously, not as evidence that blinding causes fragility, but as a signal that trial design, endpoint ascertainment, reporting practices, and patient-retention procedures may influence the relationship between missing follow-up and statistical robustness. These implications must be balanced against several limitations, including reliance on reconstructed rather than original patient-level data, incomplete and non-standardized WCLFU reporting, outcome-level non-independence when multiple endpoints or reports came from the same trial, and the substantial heterogeneity of breast cancer subtypes and treatment strategies. SIFI and SFQ should therefore be regarded as complementary evidence-appraisal metrics rather than substitutes for clinical judgment, effect size, safety, biological rationale, or guideline methodology.

In conclusion, this meta-epidemiologic analysis shows that survival outcomes from contemporary phase III breast cancer trials differ substantially in statistical robustness, even when interpreted through similar conventional measures such as hazard ratios, confidence intervals, and P values. SIFI and SFQ provide complementary perspectives by quantifying the absolute and proportional patient-level margins required to alter statistical significance interpretation, while WCLFU places these margins in the practical context of follow-up completeness. These metrics should not be used to invalidate trial conclusions or replace established assessments of efficacy, safety, endpoint relevance, and clinical value. Rather, they offer an additional layer of transparency for interpreting pivotal evidence, particularly when treatment effects are modest, P values are close to significance thresholds, follow-up is incomplete, or trial results are being considered for regulatory approval, indication expansion, or guideline adoption. Ultimately, incorporating SIFI, SFQ, and WCLFU into the appraisal of phase III breast cancer trials may help clinicians, investigators, regulators, and guideline panels distinguish statistical significance from evidential stability, thereby supporting a more transparent and context-sensitive interpretation of practice-changing breast cancer evidence.

Abbreviations

ADC	antibody-drug conjugate
AKT	protein kinase B pathway
ASCO	American Society of Clinical Oncology
CI	confidence interval
DFS	disease-free survival
EFS	event-free survival
FDA	U.S. Food and Drug Administration
HER2	human epidermal growth factor receptor 2
HR	hazard ratio
IDFS	invasive disease-free survival
IIT	investigator-initiated trial
IND	

industry-sponsored investigational new drug trial
IQR
interquartile range
KM
Kaplan–Meier
OR
odds ratio
OS
overall survival
PFS
progression-free survival
PRISMA
Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROSPERO
International Prospective Register of Systematic Reviews
RCT
randomized controlled trial
SFQ
survival fragility quotient
SIFI
survival-inferred fragility index
TNBC
triple-negative breast cancer
WCLFU
withdrawal of consent or loss to follow-up.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

All data generated or analysed during this study are included in this published article [and its supplementary information files].

Conflict of interest statement

The authors declare no competing interests.

Funding

China Academy of Medical Sciences Clinical and Translational Medicine Research Special Project 2024-I2M-C&T-C-009; National Natural Science Foundation of China (82303955); Cooperative Fund of China Academy of Medical Sciences Cancer Hospital CFA202503013

Author contributions

Y.H., D.L. and X.Z. were responsible for the initial draft writing and the data analysis-related work. S.W. and C.W. were in charge of the data collection. J.Y., X.Y. and L.G. checked the manuscript. H.H. and N.L. provided guidance for the research on the manuscript.

Acknowledgements

None.

References

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, Jemal A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2024;74(3):229–63.
2. Al Sukhun S, Koczwara B, Temin S, et al. Systemic treatment of patients with metastatic breast cancer: ASCO resource-stratified guideline. *JCO Glob Oncol.* 2024;10:e2300285.
3. Xiong X, Zheng LW, Ding Y, et al. Breast cancer: pathogenesis and treatments. *Signal Transduct Target Ther.* 2025;10(1):49. Published 2025 Feb 19.
4. Adjuvant and neoadjuvant breast cancer treatments. A systematic review of their effects on mortality.
5. U.S. Food and Drug Administration. Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics. Guidance for Industry. Published December; 2018.
6. Wheaton L, Bujkiewicz S. Use of surrogate endpoints in health technology assessment: a review of selected NICE technology appraisals in oncology. *Int J Technol Assess Health Care.* 2025;41(1):e11.
7. Jommi C, Bonfanti M, Guardigni M, et al. Pivotal Studies for Drugs About to Be Launched for Rare Diseases: Will They Better Support Health Technology Assessment and Market Access than in the Past? *J Mark Access Health Policy.* 2025;13(3):37. Published 2025 Jul 25.
8. Burstein HJ, DeMichele A, Fallowfield L, Somerfield MR, Henry NL. Endocrine and targeted therapy for hormone receptor-positive, human epidermal growth factor receptor 2-negative metastatic breast cancer—capivasertib-fulvestrant: ASCO rapid recommendation update. *J Clin Oncol.* 2024;42(12):1450–3.

9. Turner NC, Oliveira M, Howell SJ, et al. Capivasertib in hormone receptor-positive advanced breast cancer. *N Engl J Med*. 2023;388(22):2058–70.
10. Modi S, Jacot W, Yamashita T, et al. Trastuzumab deruxtecan in previously treated HER2-low advanced breast cancer. *N Engl J Med*. 2022;387(1):9–20.
11. Schmid P, Cortes J, Dent R, et al. Overall survival with pembrolizumab in early-stage triple-negative breast cancer. *N Engl J Med*. 2024;391(21):1981–91.
12. Wasserstein RL, Lazar NA. The ASA statement on p-values: context, process, and purpose. *Am Stat*. 2016;70(2):129–33.
13. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond $p < 0.05$. *Am Stat*. 2019;73(sup1):1–19.
14. 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. *J Clin Epidemiol*. 2014;67(6):622–8.
15. Garcia MVF, Duong A, Karmakar MK. Fragility index and fragility quotient in randomized clinical trials. *BJA Educ*. 2023;23(7):257–61.
16. Gonzalez MR, Acosta JI, Larios F, et al. Reverse Fragility Index: Comparing Revision Rates Between Direct Anterior and Other Approaches in Total Hip Arthroplasty. A Systematic Review of Randomized Controlled Trials. *J Arthroplasty*. 2024;39(7):1888–93.
17. 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*. 2020;3(10):e2017675.
18. Sherry AD, Liu Y, Msaouel P, et al. Survival-inferred fragility of statistical significance in phase III oncology trials. *NPJ Precis Oncol*. 2025;9:256.
19. Del Paggio JC, Tannock IF. The fragility of phase 3 trials supporting FDA-approved anticancer medicines: a retrospective analysis. *Lancet Oncol*. 2019;20(8):1065–9.
20. Olsen HE, Kao PC, Richmond C, Shulman DS, London WB, DuBois SG. Statistical Fragility of Findings From Randomized Phase 3 Trials in Pediatric Oncology. *Cancer Med*. 2024;13(24):e70356.
21. Turner NC, Oliveira M, Howell SJ, et al. Capivasertib in Hormone Receptor-Positive Advanced Breast Cancer. *N Engl J Med*. 2023;388(22):2058–70.
22. Modi S, Jacot W, Yamashita T, et al. Trastuzumab Deruxtecan in Previously Treated HER2-Low Advanced Breast Cancer. *N Engl J Med*. 2022;387(1):9–20.
23. Pocock SJ, Clayton TC, Altman DG. Survival plots of time-to-event outcomes in clinical trials: good practice and pitfalls. *Lancet*. 2002;359(9318):1686–9.
24. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71.
25. Guyot P, Ades AE, Ouwens MJNM, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Med Res Methodol*. 2012;12:9.

Figures

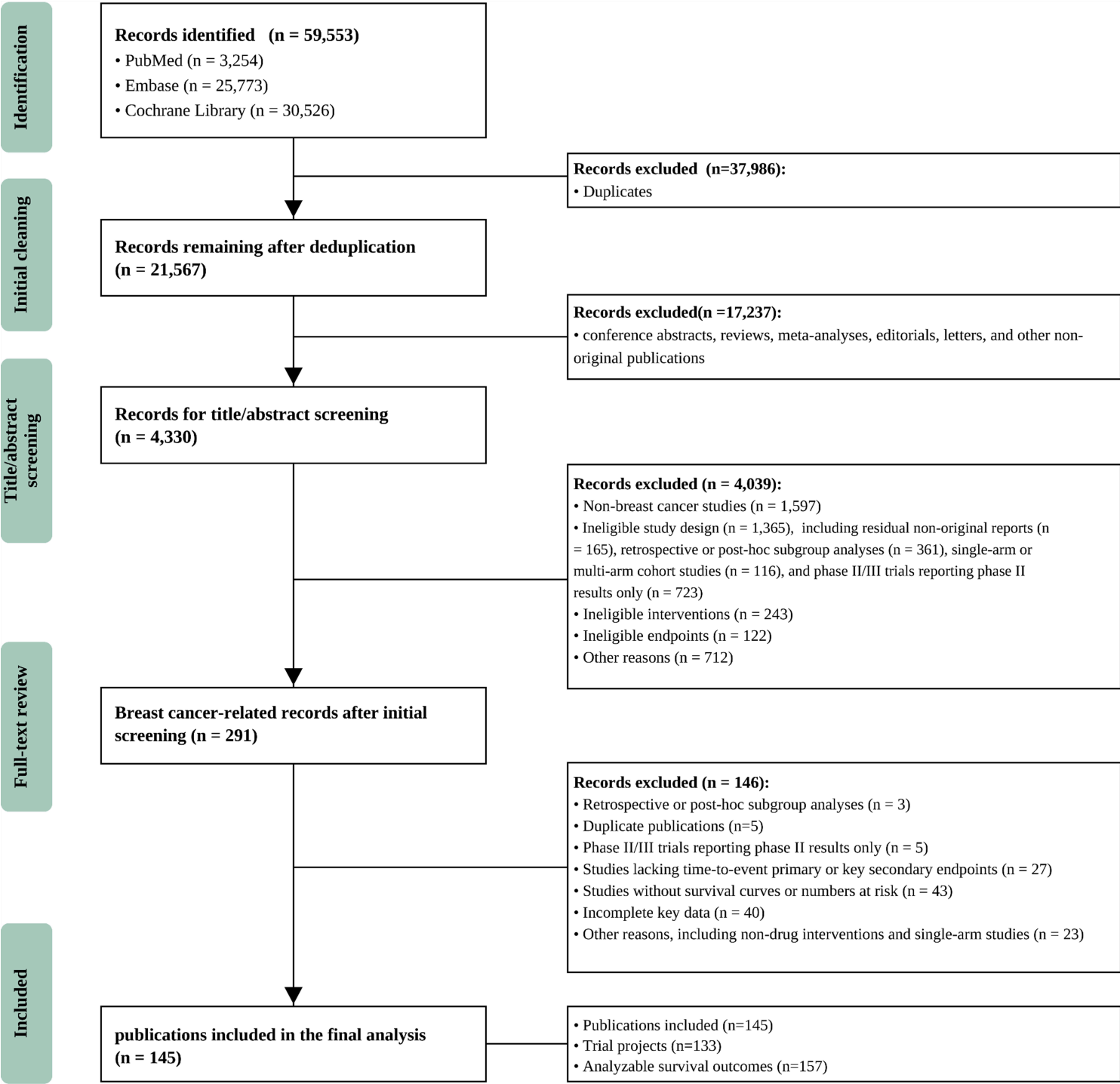


Figure 1

PRISMA flow diagram of study selection and analytical-unit construction.

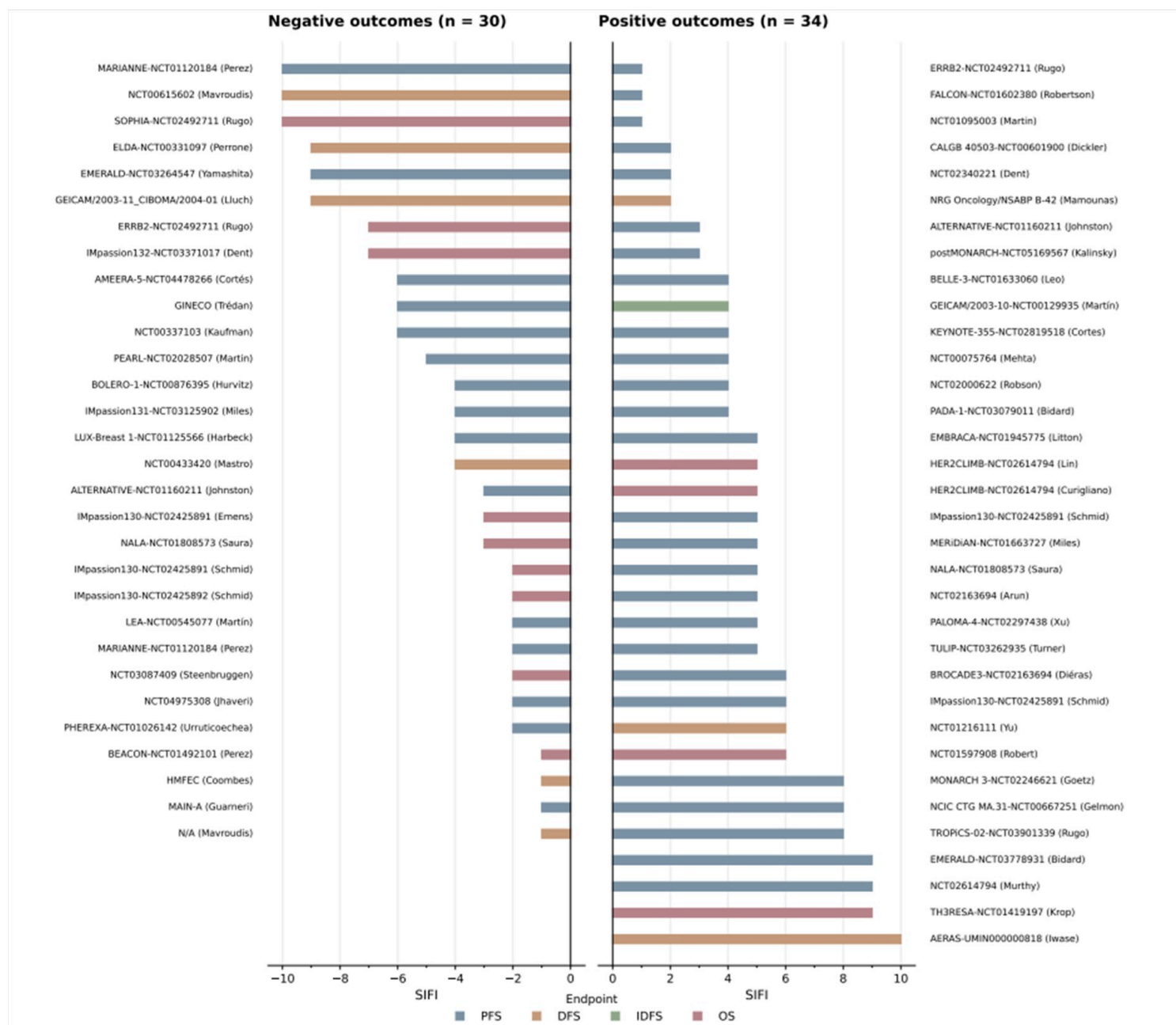


Figure 2

Distribution of outcome-level SIFI among endpoints. Note. Only endpoints with $|SIFI| \leq 10$ are shown.

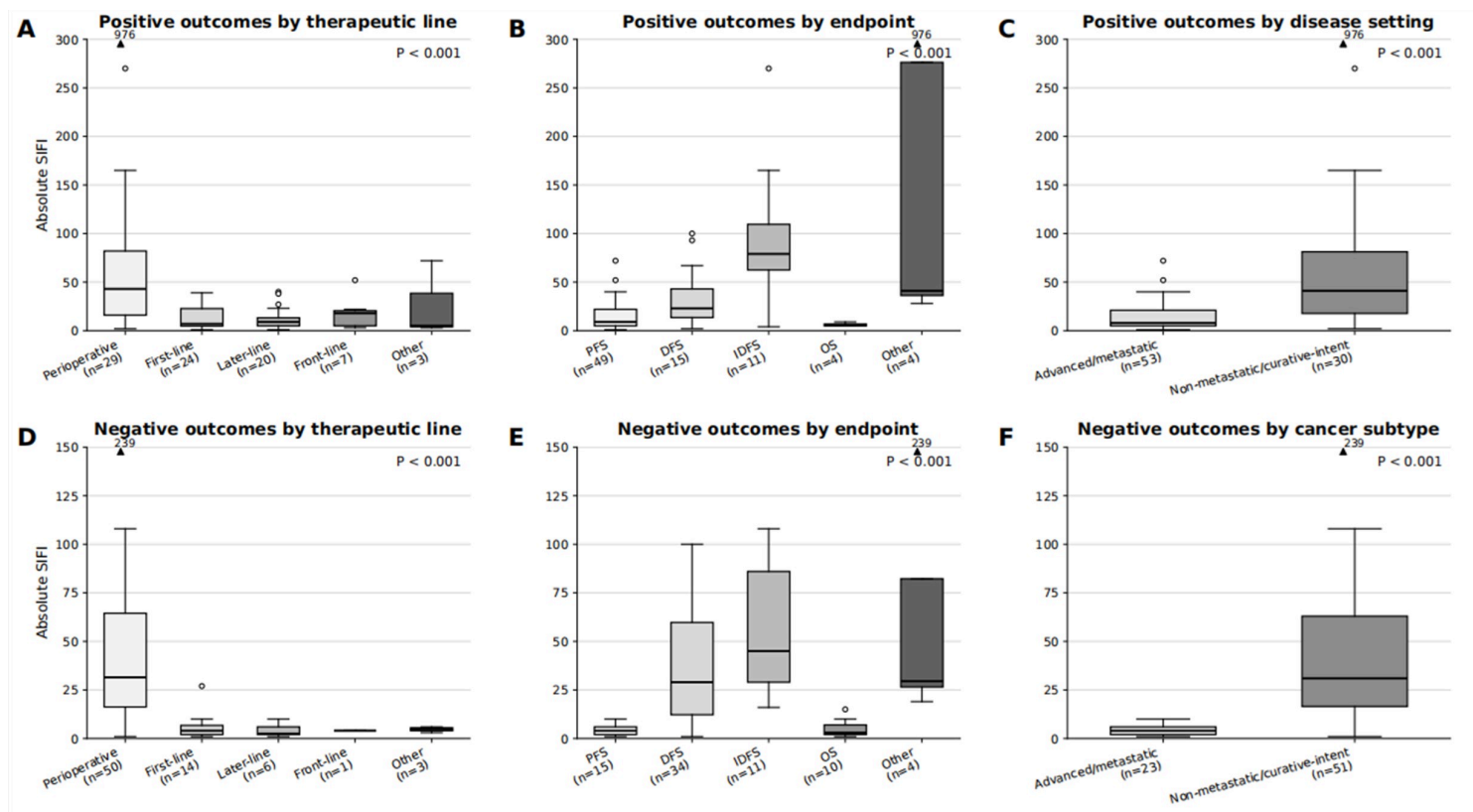


Figure 3

Key subgroup differences in absolute SIFI. Note. Boxplots show absolute SIFI by therapeutic line, endpoint type, and disease setting among positive outcomes (A–C) and negative outcomes (D–F).

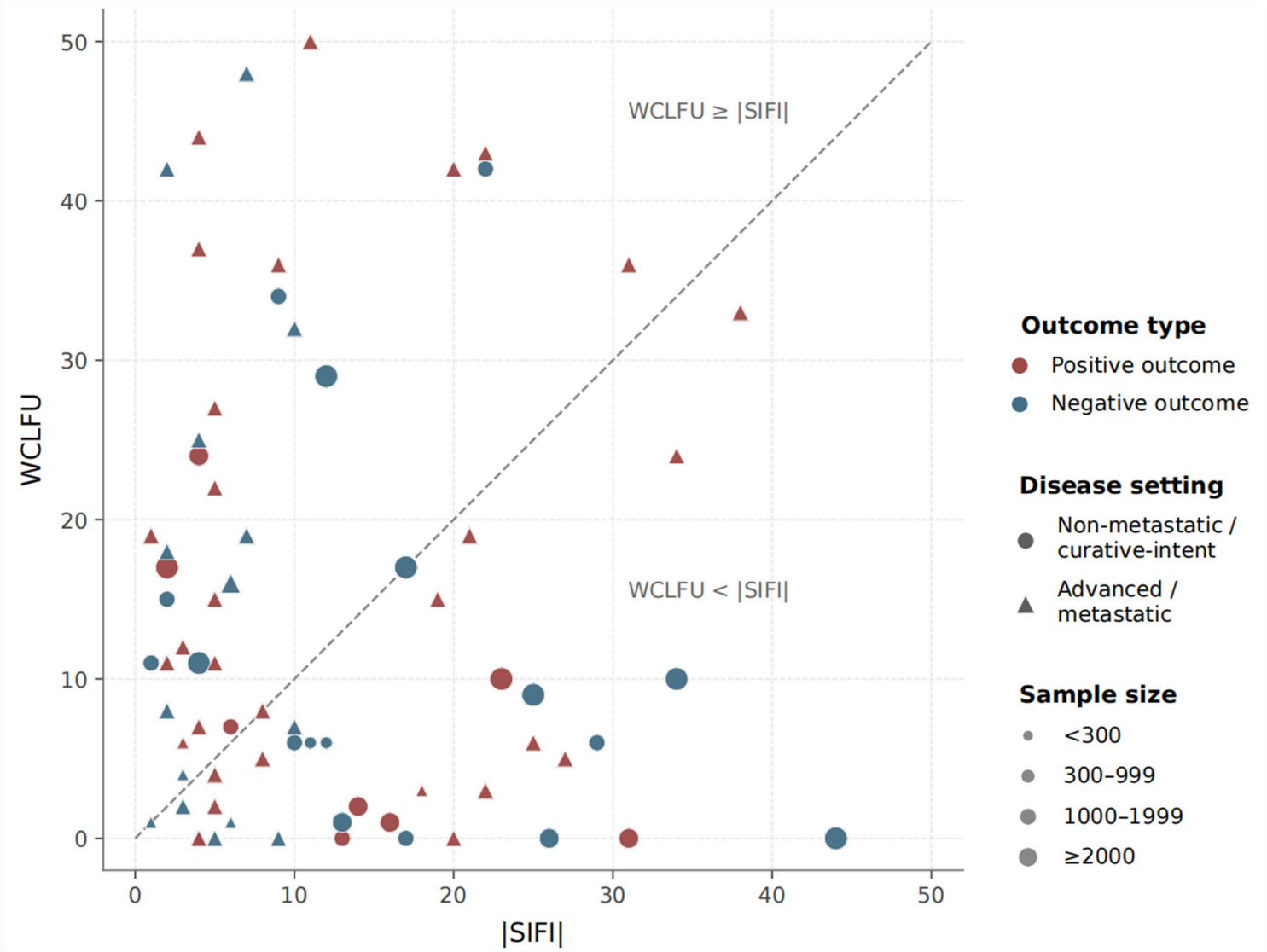


Figure 4

WCLFU relative to the absolute SIFI threshold.note:|SIFI| ≤ 50 and WCLFU ≤ 50,Outcomes outside the displayed range are not shown for readability

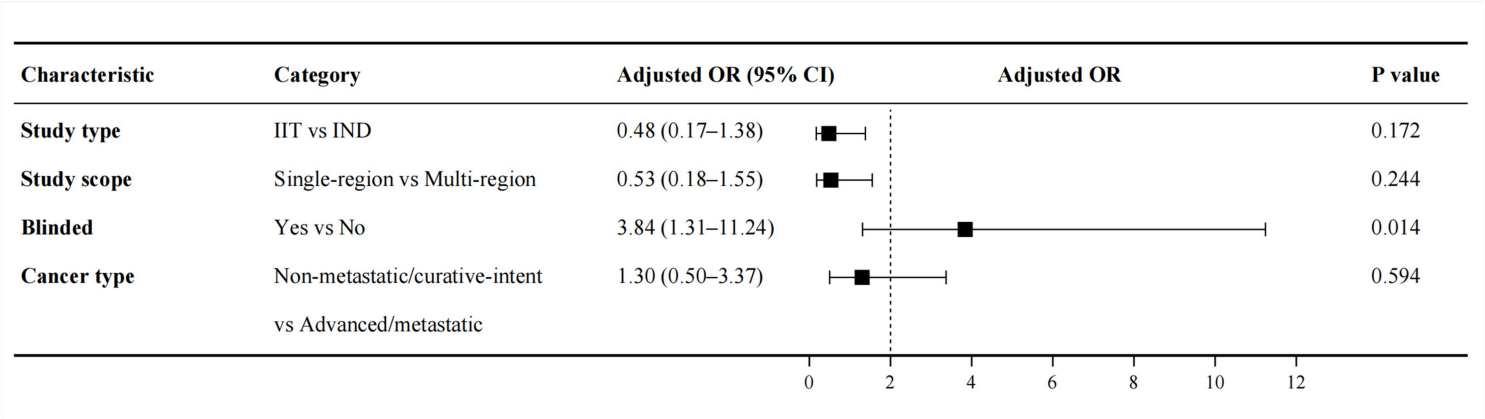


Figure 5

Core factors associated with $WCLFU \geq |SIFI|$.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryAppendix.docx](#)
- [SupplementaryTables.docx](#)
- [SupplementaryFigure1..pptx](#)