Underperformance of Contemporary Phase III Oncology Trials and Strategies for Improvement

Changyu Shen, PhD[1,2]; Enrico G. Ferro, MD[2,3]; Huiping Xu, PhD[4]; Daniel B. Kramer, MD[1,2]; Rushad Patell, MD[2,5]; and Dhruv S. Kazi, MD, MSc, MS[1,2]

ABSTRACT

Background: Statistical testing in phase III clinical trials is subject to chance errors, which can lead to false conclusions with substantial clinical and economic consequences for patients and society. **Methods:** We collected summary data for the primary endpoints of overall survival (OS) and progression-related survival (PRS) (eg, time to other type of event) for industry-sponsored, randomized, phase III superiority oncology trials from 2008 through 2017. Using an empirical Bayes methodology, we estimated the number of false-positive and false-negative errors in these trials and the errors under alternative *P* value thresholds and/or sample sizes. **Results:** We analyzed 187 OS and 216 PRS endpoints from 362 trials. Among 56 OS endpoints that achieved statistical significance, the true efficacy of experimental therapies failed to reach the projected effect size in 33 cases (58.4% false-positives). Among 131 OS endpoints that did not achieve statistical significance, the true efficacy of experimental therapies reached the projected effect size in 1 case (0.9% false-negatives). For PRS endpoints, there were 34 (24.5%) false-positives and 3 (4.2%) false-negatives. Applying an alternative *P* value threshold and/or sample size could reduce false-positive errors and slightly increase false-negative errors. **Conclusions:** Current statistical approaches detect almost all truly effective oncologic therapies studied in phase III trials, but they generate many false-positives. Adjusting testing procedures in phase III trials is numerically favorable but practically infeasible. The root of the problem is the large number of ineffective therapies being studied in phase III trials. Innovative strategies are needed to efficiently identify which new therapies merit phase III testing.

J Natl Compr Canc Netw 2021;19(9):1072–1078
doi: 10.6004/jnccn.2020.7690

[1]Richard A. and Susan F. Smith Cancer Center for Outcomes Research in Cardiology, Division of Cardiology, Department of Medicine, Beth Israel Deaconess Medical Center, [2]Harvard Medical School, and [3]Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts; [4]Department of Biostatistics, School of Medicine, Richard M. Fairbanks School of Public Health, Indiana University, Indianapolis, Indiana; and [5]Division of Hematology-Oncology, Beth Israel Deaconess Medical Center, Boston, Massachusetts.

Background

Accelerated innovation in oncologic therapies is fundamental to improving survival and quality of life for patients with cancer. Between 2011 and 2016, 68 novel oncologic drugs launched globally, and in 2018 alone, 39 new oncologic drugs were approved globally, a rate of more than 1 new drug every 2 weeks.[1] Before approval, all new therapies are tested in rigorous clinical trials to demonstrate a satisfactory risk/benefit profile, as judged by regulatory agencies. Thus, phase III trials provide crucial and important layers of evidence supporting regulatory and reimbursement decisions.

Several studies have raised concerns about whether new therapies that achieve statistical significance in phase III oncology trials offer a true improvement in clinical outcomes. Many FDA-approved oncologic therapies would not have reached statistical significance in their efficacy endpoints if a small number of trial events had been changed to nonevents (or vice versa).[2] Among phase III trials of oncologic immunotherapies with statistical significance published between 2005 and 2015, the observed effect did not reach the hypothesized effect size specified in sample size calculations for 53% of the overall survival (OS) and 38% of the progression-free survival (PFS) endpoints.[3] Among therapies approved for solid tumors between 2002 and 2014, the median OS and PFS improvements were 2.1 and 2.5 months, respectively.[4] These studies raise concerns about the number of false-positive errors, which we define in this study as new therapies that demonstrate statistically significant benefit but do not actually achieve the effect size that the trials were powered to detect (ie, the target effect size).

At the same time, other studies have also raised concerns about the amount of false-negative errors, which we define in this study as new therapies that actually achieve the target effect size but fail to demonstrate statistically significant benefit. For example, a study using clinical trial data from 2000 to 2015 showed that the overall success rate of phase III oncology trials was 48.5%, which seems dispropor-

See JNCCN.org for supplemental online content.

tionately lower than the 70% success rate of phase III trials of all other therapeutic areas combined.[5] This could be due, in part, to false-negative trials. Both false-positive and false-negative errors can have devastating consequences for patients and economic repercussions for the healthcare system. In addition, true-negative trials are also burdensome: although they often are necessary by-products of the evidence generation process and do not affect patients in the long term, a high number of true-negative trials represent a waste of societal resources that could be invested elsewhere.

Therefore, reducing the number of false-positive, false-negative, and true-negative phase III oncology trials is an urgent priority. Their relative burden in the context of current statistical methods, however, is unknown. Understanding the performance of statistical testing is fundamental to developing targeted statistical innovation to decrease the amount of all 3 undesirable outcomes and reduce patient harm and unnecessary financial expenditures. To address this gap in knowledge, we analyzed industry-sponsored phase III randomized superiority trials in oncology between 2008 and 2017 to (1) quantify the number of true-positive, false-positive, true-negative, and false-negative endpoints, and (2) determine whether alternative *P* value thresholds and/or sample size adjustments can improve the performance of phase III trials.

Methods

Selection of Trials

We identified eligible trials on ClinicalTrials.gov, because Section 801 of the FDA Amendments Act mandates registration for all phase III clinical trials of drugs and biologics initiated after September 2007 or ongoing as of December 2007.[6,7] We chose 2017 as the end year so that investigators had sufficient time to publish trial results. Trials were included if they were randomized phase III interventional superiority studies of oncologic therapies with at least one site in the United States (which ensures trial sponsor registration at ClinicalTrials.gov). We restricted our study to industry-sponsored trials because most oncologic therapies are ultimately tested via industry sponsorship, and industry sponsors tend to be more compliant with registration on ClinicalTrials.gov.[8,9]

We excluded trials that did not include OS or time to nondeath events, alone or composited with death (eg, disease progression alone, disease progression, or death), as a primary efficacy endpoint. For simplicity, we call the latter type of endpoint the “progression-related survival” (PRS). We also excluded trials testing interventions for nontreatment or noninferiority purposes (the effect size is different in superiority vs noninferiority trials), terminating early for reasons unrelated to efficacy, or with sample sizes <100 patients (Figure 1; see also supplemental eAppendix 1, available with this article at JNCCN.org).

Collection of Trial Results and Efficacy Endpoints

For each trial, we collected information on all primary efficacy endpoints of all comparisons. More details are provided in supplemental eAppendix 1. We focused our analysis on the primary efficacy endpoint of OS or PRS. The observed hazard ratio (HR) in a randomized trial is an imperfect estimate of the true unobserved HR,[10] which

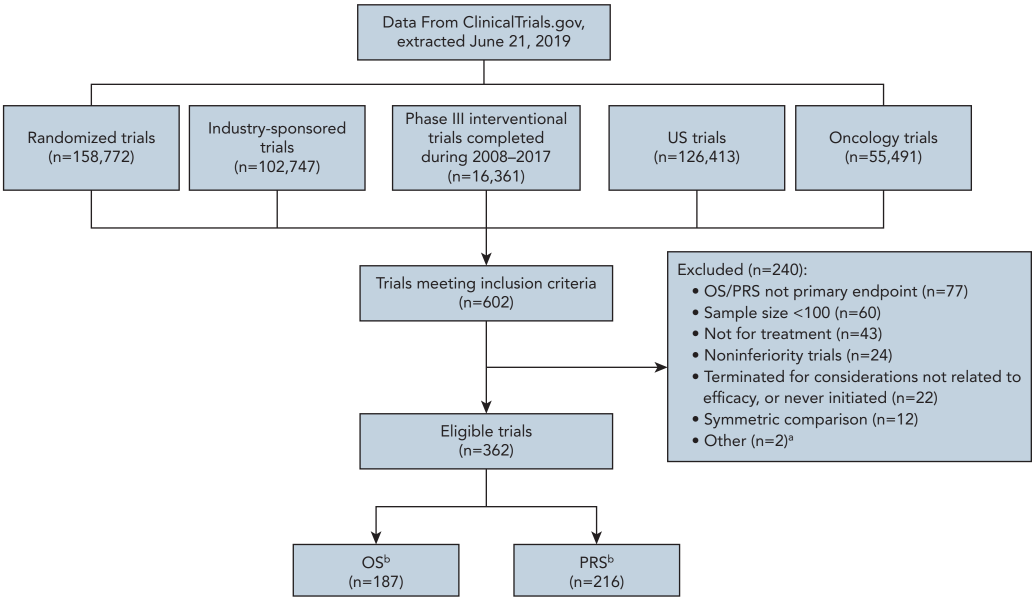
Data From ClinicalTrials.gov, extracted June 21, 2019
Randomized trials (n=158,772)
Industry-sponsored trials (n=102,747)
Phase III interventional trials completed during 2008–2017 (n=16,361)
US trials (n=126,413)
Oncology trials (n=55,491)
Trials meeting inclusion criteria (n=602)
Excluded (n=240):
• OS/PRS not primary endpoint (n=77)
• Sample size <100 (n=60)
• Not for treatment (n=43)
• Noninferiority trials (n=24)
• Terminated for considerations not related to efficacy, or never initiated (n=22)
• Symmetric comparison (n=12)
• Other (n=2)a
Eligible trials (n=362)
OSb (n=187)
PRSb (n=216)

Figure 1. Trial selection process.
Abbreviations: OS, overall survival; PRS, progression-related survival.
[a]Single-arm trial (n=1) and >90% crossover (n=1).
[b]Some trials have both OS and PRS as primary endpoints.

would be the HR of a hypothetical trial that could randomize the entire population of patients with a particular disease. We considered 3 efficacy categories based on the true unobserved HR, which in this paper is defined as control over the experimental therapy as follows:

1. Null/Negative: $\text{HR} \leq 1$ (experimental therapy has harm or no benefit relative to the control)
2. Insufficient efficacy: $\text{HR} > 1$ but less than the HR that the trial was powered to detect (the target effect size)
3. Sufficient efficacy: HR equal to or greater than the target effect size

We recognize that the target effect size in industry-sponsored trials could simply be the smallest effect that is feasible to detect. Therefore, it may not be the same as the minimum clinically important difference. Nonetheless, the target effect size is still a meaningful threshold for sufficiency because it is accepted under the oversight of the FDA and other regulatory authorities.[11]

Statistical Methods

Descriptive statistics were used to summarize trial characteristics for OS and PRS endpoints separately. For a given trial endpoint, we calculated the Z value by dividing the point estimate of the HR on a logarithmic scale by the corresponding standard error. For each trial endpoint, we then calculated the posterior probability of the 3 efficacy categories, given its Z value (or, equivalently, the P value) using an empirical Bayes method.[12–14] More details are provided in supplemental eAppendix 1. We calculated the expected count of each of the 3 efficacy categories for a group of endpoints by summing the posterior probability of each efficacy category. For instance, we can compute the expected count of endpoints with sufficient efficacy for those with $P < .05$ by summing their posterior probabilities of sufficient efficacy. We also performed similar calculations under alternative sample sizes and/or P value thresholds (supplemental eAppendix 1).

We could not find the necessary information to compute the Z values for 6.4% and 1.4% of the OS and PRS endpoints, respectively. Because the unavailability of this information was most likely due to trial sponsors not publishing negative trial results,[15] we assumed that the missing Z values have the same distribution as the observed Z values that are negative (ie, HR favors control arm). A weighting approach was used to account for this missing data mechanism when estimating the distribution of the standardized effect size (supplemental eAppendix 1).

Results

Trial Characteristics

We identified 362 eligible trials with 187 OS and 216 PRS endpoints (Figure 1). Trial characteristics for each of the 2 endpoints are summarized in Table 1. Most trials were on lung, breast, gastrointestinal, and hematologic cancers. The trials are predominately 2-arm studies of an intervention drug compared with a control treatment.

False-Positive and False-Negative Errors

Figure 2 shows the probability of each of the 3 efficacy categories for OS and PRS endpoints. For example, for an OS endpoint with $P = .05$ favoring the experimental therapy, the probability that the experimental therapy has null/negative, insufficient, or sufficient efficacy is 3.3% (blue segment), 88.6% (red segment), and 8.1% (green segment), respectively (Figure 2A). In other words, for every 100 OS endpoints with $P = .05$, experimental therapies have null/negative efficacy in approximately 3 cases, insufficient efficacy in 89 cases, and sufficient efficacy in

Table 1. Trial Characteristics, by Endpoint

Characteristic	OS n (%)	PRS n (%)	Total n (%)
Total, n	187	216	403
Tumor type			
Hematology	13 (7.0)	41 (19.0)	54 (13.4)
Breast	7 (3.7)	47 (21.8)	54 (13.4)
Gastrointestinal tract	50 (26.7)	17 (7.9)	67 (16.6)
Kidney	5 (2.7)	12 (5.6)	17 (4.2)
Lung	47 (25.1)	29 (13.4)	76 (18.9)
Ovary	1 (0.5)	18 (8.3)	19 (4.7)
Prostate	29 (15.5)	12 (5.6)	41 (10.2)
Skin	13 (7.0)	17 (7.9)	30 (7.4)
Other	22 (11.8)	23 (10.6)	45 (11.1)
Intervention type			
Drug	151 (80.8)	199 (92.1)	350 (86.9)
Biologic	35 (18.7)	16 (7.4)	51 (12.7)
Device	1 (0.5)	1 (0.5)	2 (0.5)
Allocation arms			
2	170 (90.9)	187 (86.6)	357 (88.6)
3	15 (8.0)	18 (8.3)	33 (8.2)
>3	2 (1.1)	11 (5.1)	13 (3.2)
Sample size, median (IQR)[a]	656 (485–853)	556 (353–778)	614 (402–829)
Year			
2008–2009	27 (14.4)	29 (13.4)	56 (13.8)
2010–2011	33 (17.6)	26 (12.0)	59 (14.6)
2012–2013	41 (21.9)	55 (25.5)	96 (23.8)
2014–2015	41 (21.9)	45 (20.8)	86 (21.3)
2016–2017	45 (24.1)	61 (28.2)	106 (26.3)

Abbreviations: IQR, interquartile range; OS, overall survival; PRS, progression-related survival.
[a] Calculations based on 390 available data points; "sample size" refers to the total number of subjects included in the comparison of 2 arms.

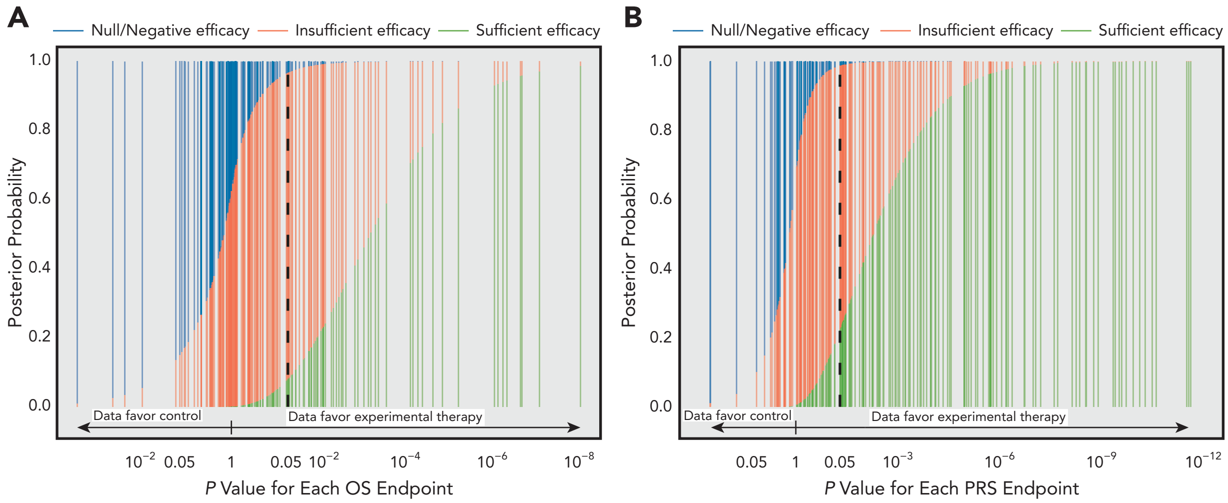

Figure 2. Probability of each efficacy category for each trial endpoint. Probability of each of the 3 efficacy categories (*y* axis) versus the 2-sided *P* value (*x* axis) for the (A) OS endpoints and (B) PRS endpoints. Each vertical line represents an endpoint from a trial, ordered by the associated *P* value. Each vertical line contains segments of different colors, and the length of these segments represents the probability that a given endpoint will fall into 1 of the 3 efficacy categories (null/negative, insufficient, and sufficient efficacy). The dashed vertical line represents the current threshold at *P*=.05. Among endpoints where data favor the experimental therapy, those with smaller *P* values have higher probability of sufficient efficacy (ie, longer green segments). Included are 175 OS and 213 PRS endpoints with efficacy results available.
Abbreviations: OS, overall survival; PRS, progression-related survival.

8 cases. Similarly, for every 100 PRS endpoints with *P*=.05, experimental therapies have null/negative efficacy in approximately 1 case, insufficient efficacy in 77 cases, and sufficient efficacy in 22 cases.

The probabilities in Figure 2 can also be viewed as an "expected count." For instance, at *P*=.05 favoring the experimental therapy, an OS endpoint contributes 0.033, 0.886, and 0.081 counts toward the categories of null/negative, insufficient, and sufficient efficacy. If we sum the lengths of all the segments by color for the endpoints on the right side of the vertical dashed line at *P*=.05, we then obtain the expected count in each efficacy category for those endpoints with statistically significant improvement. The same calculations can be performed for the endpoints on the left side of *P*=.05 to compute the expected counts in each efficacy category for those endpoints without statistically significant improvement. These counts and percentages are summarized in Table 2. The standard errors for the estimates in Table 2 are summarized in supplemental eTable 2.

The first section of Table 2 shows that among 187 OS endpoints, 56 (29.9%) demonstrated statistically significant improvement. Among these 56 OS endpoints, the experimental therapies had null/negative, insufficient, and sufficient efficacy in 0.5 (0.9%), 32.2 (57.5%), and 23.3 (41.6%) cases, respectively. Therefore, although essentially all experimental therapies from positive trials had some efficacy, more than half (ie, 0.9% + 57.5%) did not reach the target effect size and are false-positives. Of the 131 OS endpoints without statistically significant improvement, 1.2 (0.9%) therapies reached the target effect size and are false-negatives. In other words, essentially all experimental therapies from negative trials are correctly identified as not having sufficient efficacy. Aggregating all OS endpoints from trials with and without statistically significant improvement, 24.5 (13.1%) of the 187 OS endpoints tested in phase III trials have sufficient efficacy.

Among 216 PRS endpoints, 139 (64.4%) demonstrated statistically significant improvement (Table 2). Of these, the new therapies had null/negative, insufficient, and sufficient efficacy in 0.2 (0.2%), 33.8 (24.3%), and 105.0 (75.5%) cases, respectively. In other words, 75% of experimental therapies from positive trials have actually reached the target effect size, making the proportion of false-positives lower (24.5%) than that of OS endpoints (58.4%). Of the 77 PRS endpoints without statistically significant improvement, 3.3 therapies (4.2%) had reached the target effect size, making the proportion of false-negatives slightly higher than that of OS endpoints.

True-Negative Endpoints

Among 131 OS endpoints without statistical significance, 49.6 (37.9%) had null/negative efficacy and 80.2 (61.2%) had insufficient efficacy (Table 2). Summed together, 129.8 of 187 OS endpoints (69.4%) are true-negatives. Of 216 PRS endpoints, 73.7 (34.1%) are true-negative results. Thus, among all 403 endpoints, 203.5 (50.5%) are true-negatives.

Application of Alternative Phase III Designs to OS Endpoints

The first section of Table 3 shows the trial breakdown based on *P*=.005, which has been proposed in recent literature as a

Table 2. Estimates of Efficacy Categories for Therapies With and Without Statistically Significant Improvement

Efficacy Category	Statistically Significant Improvement[a]	No Statistically Significant Improvement[a]	n (%)
Overall survival			
Null/Negative	0.5 (0.9%)[b]	49.6 (37.9%)[c]	50.1 (26.8%)
Insufficient	32.2 (57.5%)[b]	80.2 (61.2%)[c]	112.4 (60.1%)
Sufficient	23.3 (41.6%)[d]	1.2 (0.9%)[e]	24.5 (13.1%)
Total	56 (100%)	131 (100%)	187 (100%)
Progression-free survival			
Null/Negative	0.2 (0.2%)[b]	23.9 (31.5%)[c]	24.1 (11.2%)
Insufficient	33.8 (24.3%)[b]	49.8 (64.3%)[c]	83.6 (38.7%)
Sufficient	105.0 (75.5%)[d]	3.3 (4.2%)[e]	108.3 (50.1%)
Total	139 (100%)	77 (100%)	216 (100%)

Testing threshold is $P<.05$.
[a] Percentages in parentheses are column percentages.
[b] False-positive.
[c] True-negative.
[d] True-positive.
[e] False-negative.

strategy to reduce false-positive errors.[16–21] When comparing the results of the first section of Table 2 with the first section in Table 3, one can see that the number of OS endpoints with statistically significant improvement decreases from 56 to 29, which is mainly due to a decrease in the number of therapies with insufficient efficacy (false-positives) from 32.2 (57.5%) to 10.1 (34.7%). Meanwhile, the number of false-negatives increases from 1.2 (0.9%) to 5.7 (3.6%).

Similarly, the subsequent sections of Table 3 show the trial breakdown resulting from a 50% sample size reduction, from a 20% sample size reduction with $P=.01$, and from a 20% sample size increase with $P=.0025$. Taken together, these data show that all these alternative testing strategies, when applied to phase III trials, reduce the power of the trials and predominately convert false-positive into true-negative endpoints.

Discussion

Our study is the first to rigorously analyze the performance of contemporary phase III oncology trials and quantify the number of false-positive, false-negative, and true-negative therapies, all of which have undesirable consequences. Our study had 4 notable findings. First, we found that the statistical testing procedures implemented in phase III oncology trials confirmed essentially all therapies with true efficacy for OS and PRS endpoints, with very few false-negatives; in other words, for all the therapies tested in phase III trials that are truly effective, essentially all of them demonstrate statistical significance. Second, we found that there is a high number of false-positives, especially for OS: among endpoints that reached statistical significance, almost 60% did not actually reach the target effect size in prolonging OS. Third, we found that 69.4% and 34.1% of OS and PRS endpoints, respectively, are true-negatives. Fourth, statistical adjustment for OS endpoints in phase III trials (such as lowering the *P* value threshold) can, in theory, reduce false-positive errors with only a small increase in false-negative errors.

The high number of false-positive trials generated by contemporary statistical testing procedures has important ramifications. First, they may result in the approval of therapies with insufficient efficacy or may even lead to worse outcomes (negative efficacy), thus exposing patients to adverse effects of interventions that are unlikely to produce meaningful health gains. From a financial perspective, a new oncologic drug costs >$100,000 USD per patient per year to payers.[22] Therefore, false-positive trials are currently producing a large financial burden for patients and payers for therapies that are ultimately ineffective.

The alternative statistical testing strategies proposed to address these problems work by decreasing power for therapies with insufficient efficacy. In so doing, they convert as many false-positives into true-negatives while tolerating the small conversion of true-positives into false-negatives, which is an inevitable by-product of this process. This procedure is numerically desirable because it improves the accuracy in statistical testing, but it has some limitations. First, these strategies can dramatically reduce the trial success rate (from 30% to between 15% and 18% for OS in our case), which can make the process practically infeasible. From an ethical perspective, the reduced power to confirm truly effective therapies can expose patients to the adverse effects of novel therapies, with a reduced likelihood of success. From a financial perspective, trial sponsors may not be willing to invest in trials with reduced power. Even if they were, our healthcare system would spend even more of the limited available resources to run negative trials. In addition, this process cannot reduce the number of ineffective therapies tested in phase III trials,

Table 3. Estimates of True Efficacy Categories for Overall Survival Endpoints, Under Alternative Testing Strategies

Sample Size	*P* Value Threshold	Efficacy Category	Statistically Significant Improvement[a]	No Statistically Significant Improvement[a]
100%	.005	Null/Negative	0.03 (0.1%)[b]	50.0 (31.6%)[c]
		Insufficient	10.1 (34.7%)[b]	102.3 (64.8%)[c]
		Sufficient	18.9 (65.2%)[d]	5.7 (3.6%)[e]
		Total	29 (100%)	158 (100%)
50%	.05	Null/Negative	0.5 (1.4%)[b]	49.6 (32.8%)[c]
		Insufficient	17.2 (47.8%)[b]	95.2 (63.1%)[c]
		Sufficient	18.3 (50.8%)[d]	6.2 (4.1%)[e]
		Total	36 (100%)	151 (100%)
80%	.01	Null/Negative	0.08 (0.2%)[b]	50.0 (31.9%)[c]
		Insufficient	11.3 (37.6%)[b]	101.1 (64.4%)[c]
		Sufficient	18.7 (62.2%)[d]	5.9 (3.7%)[e]
		Total	30 (100%)	157 (100%)
120%	.0025	Null/Negative	0.01 (0.03%)[b]	50.1 (31.9%)[c]
		Insufficient	10.0 (33.3%)[b]	102.4 (65.2%)[c]
		Sufficient	20.0 (66.6%)[d]	4.5 (2.9%)[e]
		Total	30 (100%)	157 (100%)

[a] Percentages in parentheses are column percentages.
[b] False-positive.
[c] True-negative.
[d] True-positive.
[e] False-negative.

which is alarmingly high in our study. Specifically, we found that the combination of false-positive and true-negative OS endpoints represents 87% of the total OS endpoints investigated over 10 years of phase III clinical trials in the United States, suggesting a very large group of ineffective therapies tested in phase III trials. Therefore, alternative testing strategies for phase III trials are unlikely to be practically helpful.

Instead of adjusting the testing strategy in phase III trials, we believe the solution lies in improving the statistical standards that determine which therapies are advanced to phase III trials to test OS endpoints. This can be achieved by reducing the number of false-positives in phase II trials. The same strategies suggested to improve phase III trials, such as adjusting the *P* value threshold and/or sample sizes, could instead be adopted in phase II trials. Importantly, the ethical weaknesses identified when applying these strategies to phase III trials could turn into potential advantages in phase II trials. Tolerating lower power in phase II trials would reduce the number of futile phase III trials. Because phase III trials are much larger than phase II trials, this strategy would effectively expose far fewer patients to therapies without sufficient efficacy, which is ethically desirable. In the present work, we did not conduct the calculations to quantify how the change in phase II trial design can reduce false-positives and true-negatives in phase III trials. Nonetheless, some evidence already demonstrates a strong correlation of PRS and OS endpoints between phase II and the corresponding phase III trials, which suggests that therapies that excel in phase II may also excel in phase III.[23] Future studies should analyze available phase II trial data to better understand whether it is feasible to alter the criteria that currently define the success of phase II trials.

There are potential limitations to our approach. First, trials terminated early have smaller sample sizes, leading to underestimation of the effect size. Thus, we could have underestimated the number of endpoints in the sufficient efficacy category. Second, trials with multiple arms and/or both PRS and OS primary endpoints could adopt more stringent thresholds than $P<.05$ to adjust for multiple comparison. This would enlarge the sample size, leading to overestimation of the effect size. Thus, we could have overestimated the number of endpoints in the sufficient efficacy category. However, sensitivity analyses showed that the impact of these 2 limitations on the results was negligible (supplemental eAppendix 1). Third, some trials are designed to achieve >80% power, leading to sample size inflation and overestimation of the trials with sufficient efficacy. Fourth, because of practical and financial considerations, the target effect size may not be the same as the minimum clinically important difference. Nonetheless, it serves as a reasonable threshold to study the problem at hand. Fifth, we excluded trials with <100 participants to ensure the validity of the statistical analysis, which could disproportionally exclude rare cancers or subsets of common malignancies. Finally, national cooperative-sponsored trials were excluded, and hence the results mainly reflect industry-sponsored trials.

Conclusions

Our analysis of the statistical performance of current phase III superiority trials in oncology shows an alarmingly high number of false-positive therapies, namely those deemed to prolong survival based on current statistical testing but that do not actually achieve the target effect size. Equally alarming, we found a high number of true-negative therapies. The large number of ineffective therapies (ie, false-positive therapies plus true-negative therapies) make the statistical adjustment in phase III trials ethically questionable for patients and financially unjustifiable for sponsors. A better solution is to apply more stringent statistical criteria to phase II trials. This strategy would increase the proportion of truly effective therapies that are advanced to phase III trials, subsequently reducing false-positives and true-negatives and improving trial success rate. Ultimately, this is the strategy that can potentially reduce unnecessary healthcare expenditures and, most important, improve patient outcomes.

Submitted September 14, 2020; final revision received November 29, 2020; accepted for publication November 30, 2021.
Published online June 21, 2021.

Author contributions: *Study design*: Shen. *Data analysis*: Shen, Xu. *Data interpretation*: Shen, Ferro. *Writing – original draft*: Shen, Ferro. *Writing – review and editing*: Shen, Ferro, Kramer, Patell, Kazi. *Critical feedback for further interpretation of results*: Shen, Ferro, Kramer, Patell, Kazi. *Final approval*: All authors.

Disclosures: The authors have disclosed that they have not received any financial consideration from any person or organization to support the preparation, analysis, results, or discussion of this article.

Correspondence: Changyu Shen, PhD, Smith Center for Outcomes Research, Department of Medicine, Beth Israel Deaconess Medical Center, 375 Longwood Avenue, 4th Floor, Boston, MA 02215.
Email: changyushen312@gmail.com

References

1. Wilson BE, Jacob S, Yap ML, et al. Estimates of global chemotherapy demands and corresponding physician workforce requirements for 2018 and 2040: a population-based study. Lancet Oncol 2019;20:769–780.
2. Del Paggio JC, Azariah B, Sullivan R, et al. Do contemporary randomized controlled trials meet ESMO thresholds for meaningful clinical benefit? Ann Oncol 2017;28:157–162.
3. Lawrence NJ, Roncolato F, Martin A, et al. Effect sizes hypothesized and observed in contemporary phase III trials of targeted and immunological therapies for advanced cancer. JNCI Cancer Spectr 2018;2:pky037.
4. Fojo T, Mailankody S, Lo A. Unintended consequences of expensive cancer therapeutics—the pursuit of marginal indications and a me-too mentality that stifles innovation and creativity: the John Conley Lecture. JAMA Otolaryngol Head Neck Surg 2014;140:1225–1236.
5. Wong CH, Siah KW, Lo AW. Estimation of clinical trial success rates and related parameters. Biostatistics 2019;20:273–286.
6. Food and Drug Administration Amendments Act of 2007. Pub L No. 110-85, § 801, 121 Stat 823.
7. Zarin DA, Tse T, Menikoff J. Federal human research oversight of clinical trials in the United States. JAMA 2014;311:960–961.
8. Anderson ML, Chiswell K, Peterson ED, et al. Compliance with results reporting at ClinicalTrials.gov. N Engl J Med 2015;372:1031–1039.
9. Anderson ML, Peterson ED. Compliance with results reporting at ClinicalTrials.gov [letter]. N Engl J Med 2015;372:2370–2371.
10. Slutsky DJ. Statistical errors in clinical studies. J Wrist Surg 2013;2:285–287.
11. Ocana A, Tannock IF. When are “positive” clinical trials in oncology truly positive? J Natl Cancer Inst 2011;103:16–20.
12. Efron B. Empirical Bayes deconvolution estimates. Biometrika 2016;103:1–20.
13. Shen C. Interval estimation of a population mean using existing knowledge or data on effect sizes. Stat Methods Med Res 2019;28:1703–1715.
14. Shen C, Li X. Using previous trial results to inform hypothesis testing of new interventions. J Biopharm Stat 2018;28:884–892.
15. Chen YP, Liu X, Lv JW, et al. Publication status of contemporary oncology randomised controlled trials worldwide. Eur J Cancer 2016;66:17–25.
16. Adibi A, Sin D, Sadatsafavi M. Lowering the P value threshold. JAMA 2019;321:1532–1533.
17. Benjamin DJ, Berger JO, Johannesson M, et al. Redefine statistical significance. Nat Hum Behav 2018;2:6–10.
18. Ioannidis JPA. The proposal to lower P value thresholds to .005. JAMA 2018;319:1429–1430.
19. McShane BB, Gal D, Gelman A, et al. Abandon statistical significance. Am Stat 2019;73(Suppl 1):235–245.
20. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond “$p < 0.05$.” Am Stat 2019;73(Suppl 1):1–19.
21. Wayant C, Scott J, Vassar M. Evaluation of lowering the P value threshold for statistical significance from .05 to .005 in previously published randomized clinical trials in major medical journals. JAMA 2018;320:1813–1815.
22. Mailankody S, Prasad V. Five years of cancer drug approvals: innovation, efficacy, and costs [letter]. JAMA Oncol 2015;1:539–540.
23. Vreman RA, Belitser SV, Mota ATM, et al. Efficacy gap between phase II and subsequent phase III studies in oncology. Br J Clin Pharmacol 2020;86:1306–1313.

See JNCCN.org for supplemental online content.

JNCCN

National Comprehensive Cancer Network®

JOURNAL OF THE NATIONAL COMPREHENSIVE CANCER NETWORK

Supplemental online content for:

Underperformance of Contemporary Phase III Oncology Trials and Strategies for Improvement

Changyu Shen, PhD; Enrico G. Ferro, MD; Huiping Xu, PhD; Daniel B. Kramer, MD; Rushad Patell, MD; and Dhruv S. Kazi, MD, MSc, MS

J Natl Compr Canc Netw 2021;19(9):1072–1078

eAppendix 1: Supplemental Materials

eAppendix 1. Supplemental Materials

Selection of Eligible Trials

We downloaded the zipped package of pipe-delimited files from the Clinical Trials Transformation Initiative (CITI) website (https://www.ctti-clinicaltrials.org/) on June 21, 2019. A trial was included if it meets all of the following inclusion criteria:

(I1) Phase III testing a medical intervention and completed during 2008–2017: restrict fields “study_type” to “Interventional”, “phase” to phase 3”, and “primary_completion_date” to January 1, 2008–December 31, 2017 using data file “studies”.
(I2) Randomized: restrict the field “allocation” to “randomized” using data file “designs”.
(I3) Industry-sponsored: restrict the field “agency_class” to “industry” using data file “sponsors”.
(I4) Inclusion of at least one site in the United States: select trials with at least one occurrence of “United States” under the field “name” using data file “countries”.
(I5) Inclusion of cancer as one target condition: select trials with field “downcase_mes_term” including at least one of the following terms: carcinoma, cancer, leukemia, lymphoma, myeloma, glioblastoma, melanoma, sarcoma, osteosarcoma, mesothelioma, tumor, neoplasm, and neoplasms using data file “browse_conditions”.

We then merge the 5 data files to generate the list of trials meeting inclusion criteria.
A trial was excluded if it meets one of the following exclusion criteria:

(E1) The primary endpoint(s) does not include overall survival (OS) or progression-related survival (PRS): manual review.
(E2) Sample size is <100: remove trials with the field “enrollment” <100, then manual review the rest.
(E3) Intervention not for treatment purpose: restrict the field “primary_purpose” to “treatment”.
(E4) Include test(s) of noninferiority: remove trials with the field “non_inferiority_type” equal to “non-inferiority” or “other”, then manual review the rest.
(E5) The trials are terminated for considerations not related to efficacy, or never initiated: remove trials with the field “overall_status” equal to “withdrawn” or “suspended”, then manual review the rest.
(E6) The comparison is symmetric with respect to the two interventions without a clear reference: manual review.
(E7) Other factors that could violate the validity of the comparison results: manual review.

We searched resources in the following order until we were able to collect trial results: research articles published in peer-reviewed scientific journals, ClinicalTrials.gov trial result databases, trial sponsors’ clinical summary reports, conference abstracts, press releases, and direct contact with trial sponsors.

Statistical Computations

Computation of the *Z* Statistic

For each $(1-\alpha)\times 100\%$ confidence interval (l, u) of the hazard ratio (HR), the Z statistic can be calculated as

$$Z=\frac{2\Phi^{-1}(1-\frac{\alpha}{2})\ln\sqrt{lu}}{\ln\sqrt{lu}}.$$

If a confidence interval is not available, the Z statistic can be calculated based on the P value of the log-rank test as

$$Z=D\Phi^{-1}\left(1-\frac{p}{2}\right),$$

where $D=1$ if HR estimate >1, and –1 otherwise.

Estimation

The standardized effect size is the HR at the logarithm scale divided by its standard error. Under the proportional hazard assumption, the standardized effect size θ can be written as

$$\theta=\sqrt{\pi(1-\pi)d}\ln\beta,$$

where π is the proportion of patients allocated to the experimental therapy arm, d is the number of events of the 2 arms combined at the time of analysis, and β is the HR (reference over intervention). Thus, larger positive value of θ indicates

(*continued on next page*)

eAppendix 1. Supplemental Materials (cont.)

stronger efficacy. As phase III trials typically involve a large number of patients, the Z statistic has an approximate normal distribution conditional on θ:

$$Z \sim N(\theta, 1).$$

In other words, Z is a noisy version of θ with θ as the mean and a unit standard deviation. The sample size calculation essentially involves choosing a value d that amplifies the hypothesized HR (at the logarithm scale) into a target value θ_0 such that Z has high probability (power) to pass the critical value. In a typical setting, the critical value is 1.96 (corresponding to type I error rate = 0.025) and the power is set at 80%, which lead to $\theta_0 = 2.8$. Clearly, $\theta \leq 0$ means null or negative efficacy. Therefore, therapies with null/negative efficacy, insufficient and sufficient efficacy correspond to $\theta \leq 0$, $0 < \theta < \theta_0$, and $\theta > \theta_0$.

Computation of the Distribution of the Standardized Effect Size, $g(\theta)$

With Z values from a large number of trials, we can infer the distribution of θ, $g(\theta)$, which is the central element of our methodology. There are parametric, semiparametric and nonparametric methods that can be used to estimate $g(\theta)$.[1–4] In this article, we will use a semiparametric method to estimate $g(\theta)$,[1] which enjoys a good balance of robustness and precision. In this method, $g(\theta)$ is assumed to be from a class of distributions in the form of cubic spline (at logarithm scale). We applied maximum likelihood estimation (MLE) with a grid precision of 0.02.

Missing Z values most likely indicate that the results were negative, which was why it was hard to find. To be conservative, we assume these values have the same distribution as those observed non-positive Z values. For a given endpoint (eg, OS), let n_M and $n_{\leq 0}$ be the numbers of missing Z values and the observed non-positive Z values, we then assign a weight 1 to all observed positive Z values and a weight of $(1 + n_M / n_{\leq 0})$ for all $n_{\leq 0}$ non-positive Z values in the MLE. This is a conservative weighting strategy assuming that the distribution of the missing Z values is the same as that of the non-positive Z values. For both OS and PFS, a 3-degree of freedom produced the best Akaike information criterion (AIC). Therefore, the results of this manuscript are based on a cubic spline with 3 degrees of freedom.

Computation of Posterior Probabilities

Once we obtain the estimate of $g(\theta)$, we can compute the posterior probabilities given Z by plugging the estimate in to the formula shown in eTable 1.

eTable 1. Posterior Probability of Each Efficacy Category at $Z = z_0$, Where $\theta_0 = 2.8$, $\phi(\cdot)$ is the Probability Density Function of the Standard Normal Distribution

Efficacy Category	Formulae
Null/Negative efficacy	$\dfrac{\int_{-\infty}^{0} g(\theta)\phi(z_0-\theta)d\theta}{\int_{-\infty}^{\infty} g(\theta)\phi(z_0-\theta)d\theta}$
Insufficient efficacy	$\dfrac{\int_{0}^{\theta_0} g(\theta)\phi(z_0-\theta)d\theta}{\int_{-\infty}^{\infty} g(\theta)\phi(z_0-\theta)d\theta}$
Sufficient efficacy	$\dfrac{\int_{\theta_0}^{\infty} g(\theta)\phi(z_0-\theta)d\theta}{\int_{-\infty}^{\infty} g(\theta)\phi(z_0-\theta)d\theta}$

Computation of the Distribution of Standardized Effect Size Under Alternative Sample Sizes

Under alterative sample size $d_A = kd(k \neq 1)$, the standardized effect size becomes $\theta_A = \sqrt{k}\theta$. Therefore, the distribution of the alternative standardized effect size is $g_A(\theta_A) = \frac{1}{\sqrt{k}} g\left(\theta_A / \sqrt{k}\right)$. The marginal probability density function of the Z statistic under the alternative sample size is $f_A(z) = \int_{-\infty}^{\infty} g_A(\theta_A)\phi(z-\theta_A)d\theta_A$.

(*continued on next page*)

eAppendix 1. Supplemental Materials (cont.)

Standard Errors of Table 2

eTable 2. Standard Errors of Quantities in Table 2

	Efficacy Category	Statistically Significant Improvement	No Statistically Significant Improvement	n (%)
Overall survival[a]	Null/Negative	0.05 (0.19%)	7.5 (5.0%)	7.5 (4.0%)
	Insufficient	3.4 (5.9%)	6.9 (5.0%)	8.3 (4.4%)
	Sufficient	4.4 (6.0%)	0.3 (0.3%)	4.6 (2.5%)
	Total	5.5 (–)	5.5 (–)	187 (100%)
Progression-free survival[a]	Null/Negative	0.02 (0.02%)	4.1 (3.7%)	4.0 (1.9%)
	Insufficient	2.2 (2.1%)	3.4 (3.3%)	5.5 (2.5%)
	Sufficient	6.4 (2.1%)	0.2 (0.5%)	6.4 (3.0%)
	Total	5.4 (–)	5.4 (–)	216 (100%)

[a] Percentages in parentheses are column percentages.

Underestimation and Overestimation of the Category of Sufficient Efficacy

Underestimation Due to Early Termination

There are 24 Z values for OS and 16 Z values for PRS endpoints that were based on trials terminated early due to interim analysis or other reasons. The absolute value of the standardized effect size of these trials is less than what would have been had they not been terminated earlier. Therefore, the threshold for sufficient efficacy at the scale of standardized effect size for these trials should be lower. This means we could have underestimated the probability of sufficient efficacy, leading to underestimate of the counts of sufficient efficacy in groups with and without statistically significant improvement in Table 2. If we assume that most trials do not start interim analysis until 25% of the events have occurred, then the standardized effect size calculated at this first interim analysis would be half of the value at the time when 100% of the planned events have occurred. Thus, the threshold of hypothesized effect size at the scale of standardized effect size should be 1.4 instead of 2.8. We recalculated the posterior probability of sufficient efficacy with the threshold of 1.4 for the 24 and 16 for OS and PFS endpoints terminated earlier, and examined the added expected counts of sufficient efficacy from these trials. For the statistically significant group, there is 0.5 and 0.8 extra expected count of the sufficient efficacy category for OS and PRS, respectively. For the statistically insignificant group, there are 1.4 and 2.5 extra expected counts of the sufficient efficacy category for OS and PRS, respectively.

Overestimation Due to Smaller Than 2.5% Type I Error Rate

Thirty-seven trials have multiple arms and/or both PFS and OS as primary endpoints and adopted a more stringent threshold than $P<.05$ to adjust for multiple comparison. Specifically, 35 trials have 2 endpoints and 2 trials have 4 points. We assumed that testing of endpoints in these trials are based on a threshold of $P<.025$ (eg, Bonferroni correction for 2 endpoints in a trial). The threshold for the target effect size at the scale of standardized effect size is 3.08. We recalculated the posterior probability of sufficient efficacy with the threshold of 3.08 for the 78 endpoints, and examined the reduced expected counts of sufficient efficacy from these trials. For the statistically significant group, there is one reduced expected count of the sufficient efficacy category for both OS and PRS. For the statistically insignificant group, there is 0.1 and 0.4 reduced expected counts of the sufficient efficacy category for OS and PRS, respectively.

References

1. Efron B. Empirical Bayes deconvolution estimates. Biometrika 2016;103:1–20.
2. Carroll RJ, Hall P. Optimal rates of convergence for deconvolving a density. J Am Stat Assoc 1988;83:1184–1186.
3. Fan J. On the optimal rates of convergence for nonparametric deconvolution problems. Ann Stat 1991;19:1257–1272.
4. Shen C. Interval estimation of a population mean using existing knowledge or data on effect sizes. Stat Methods Med Res 2019;28:1703–1715.