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for considerations not related to efficacy, or never initiated: remove trials with the field "overall_status" equal to "withdrawn" or "suspended", then manual review the rest.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

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

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	$\frac{\int_{-\infty}^{0} g(\theta)\phi(z_0-\theta)d\theta}{\int_{-\infty}^{\infty} g(\theta)\phi(z_0-\theta)d\theta}$
Insufficient efficacy	$\frac{\int_{0}^{\theta_0} g(\theta)\phi(z_0-\theta)d\theta}{\int_{-\infty}^{\infty} g(\theta)\phi(z_0-\theta)d\theta}$
Sufficient efficacy	$\frac{\int_{\theta_0}^{\infty} g(\theta)\phi(z_0-\theta)d\theta}{\int_{-\infty}^{\infty} g(\theta)\phi(z_0-\theta)d\theta}$

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.