Clinical Versus Statistical Significance in Studies of Thoracic Malignancies

Suzanne E. Dahlberg, PhD,[a,*] Edward L. Korn, PhD,[b] Jennifer Le-Rademacher, PhD,[c] Sumithra J. Mandrekar, PhD[c]

[a]*Department of Pediatrics, Boston Children's Hospital, Boston, Massachusetts*
[b]*National Cancer Institute, Bethesda, Maryland*
[c]*Mayo Clinic, Rochester, Minnesota*

Received 1 April 2020; revised 3 June 2020; accepted 11 June 2020
Available online - 21 June 2020

The scientific synergy between statistics and medical research often leads to controversy when considering statistical versus clinical significance of findings from studies that are neither overwhelmingly practice-changing nor clearly negative. Studies can reach statistical significance but provide evidence that is not clinically meaningful, or results could not be statistically significant but very clinically relevant. To fully appreciate the debate about studies that fall in this gray area, one must understand how to interpret several features of statistical design and the interpretation of results.

The conduct of clinical trials dictates that a rigorous statistical design a priori targets an effect size, which can be thought of as the true difference in outcome that a study hopes to detect when comparing two regimens with one another or when comparing a single regimen to a reliable historical control. Statistical inference, based on the distributional assumptions of the design, is used to formally test the prestated hypothesis about the effect size resulting in p values, which assist in the determination of a study's success but do not provide any information regarding the magnitude of the observed effect size in the trial. The p values indicate statistical but not clinical significance; they are the metric for the determination of study success or failure after they are compared with the type I error rate, the probability of observing a false-positive result, allocated in the study design. A threshold of 0.05 is thought to be the conventional type I error rate; but in fact, the origin of this threshold is arbitrary, and in practice designs, often have lower or higher false-positive thresholds depending on design features such as adjustments for multiple comparisons or phase of development, respectively. For example, phase 3 studies often test coprimary end points resulting in type I error rates of less than 0.05 in the design, whereas phase 2 trials often relax type I error rates to as high as 0.10 or 0.20.[1] The p values should be interpreted as the probability that the study results (as good as or better than observed) occurred by chance when the null hypothesis is true.

Clinical significance is far more subjective than statistical significance. Clinical significance is defined by many parameters, including the observed effect size, primary end point, safety profile, financial toxicity, quality of life, availability of a companion diagnostic for identification of patients likely to benefit the most, demographics of the enrolled population, treatment adherence, crossover, and many others. The past 15 years of oncology drug development have revealed monumental success in bringing forth incredibly effective targeted therapies and immunotherapy for patients with lung cancer; therefore, expectations are higher than novel therapies will result in changes to current practice and substantial benefits to patient outcomes. These outcomes could be associated solely with efficacy assessments but also could be evaluated through relative trade-offs in adverse events, cost-effectiveness, health care convenience (e.g., oral versus intravenous administration), quality of life, and others. With the lack of dramatic treatment effects, incremental advancements use precious resources and time without much

*Corresponding author.

Disclosure: The authors declare no conflict of interest.

Address for correspondence: Suzanne E. Dahlberg, PhD, Boston Children's Hospital, 300 Longwood Ave., Mailstop BCH 3200, Boston, MA 02115. E-mail: suzanne.dahlberg@childrens.harvard.edu

ISSN: 1556-0864

https://doi.org/10.1016/j.jtho.2020.06.007

benefit to patients. Statistics can be used to assist in the decision-making process to objectively define clinical relevance.

As clinicians acknowledge the heterogeneity of patients and their cancers, in the same way statisticians study the variability that results from that heterogeneity. There are always patients who respond exceptionally well and patients who, unfortunately, respond more poorly than most trial participants; but the most effective therapies will exhibit a treatment effect despite that variability. Confidence intervals (CIs), when presented in tandem with p values, can help provide clarity around the determination of clinical significance because they provide the range of possible values consistent with the size of the observed effect. The benefit of CIs is that they provide a clinical context in the measurement of the direction of a treatment effect, unlike a p value. For example, the KEYNOTE-010 trial[2] studied two doses of pembrolizumab versus docetaxel in previously treated patients with NSCLC who were positive for programmed death-ligand 1 and was designed with coprimary end points of overall survival (OS) and progression-free survival (PFS). The results for the latter end point were not statistically significant because the observed p values for the PFS comparisons were not lower than the predefined threshold of p value less than 0.001. The observed hazard ratio (HR) for PFS for pembrolizumab 2 mg/kg versus docetaxel was 0.88 (95% CI: 0.74–1.05, $p = 0.07$); for pembrolizumab 10 mg/kg versus docetaxel, HR was 0.79 (95% CI: 0.66–0.94, $p = 0.004$). We raise this as an example of a study in which a p value is less than the conventional p value of 0.05 and is not statistically significant per the statistical design; however, considering the CIs for the PFS HRs, it is clear that the direction of the benefit is mostly, but not entirely, below a value of 1 (trending toward clinical benefit). OS clearly met the predefined criteria for statistical significance (one-sided $p < 0.00825$), and the upper bound of the CIs are estimated well below the value of one; for pembrolizumab 2 mg/kg versus docetaxel HR was 0.71 (95% CI: 0.58–0.88, $p = 0.0008$); for pembrolizumab 10 mg/kg versus docetaxel HR was 0.61 (95% CI: 0.49–0.75, $p < 0.0001$). We would typically recommend that the confidence level of a CI match the design-specified significance level for the end point, which does not seem to have been done in the reporting of this trial; however, the statistical concept about the general direction of the results still applies. For the dose approved by the Food and Drug Administration (2 mg/kg), the observed result (HR $=$ 0.71) is statistically consistent with an HR as promising as 0.58 and not worse than 0.88 based on the corresponding CI, clearly revealing statistical significance for efficacy and establishing clinical significance.

An example of a study with statistically significant but arguably clinically insignificant results is the REVEL trial,[3] which randomized patients with metastatic NSCLC to ramucirumab plus docetaxel versus placebo plus docetaxel as second-line treatment after disease progression on platinum-based therapy. From the time of initiation, the study was designed to detect a modest effect size, an OS HR of 0.816, with 85% power while controlling the two-sided type I error rate at the 0.05 level. The study was, therefore, very large, having randomized 1253 patients and resulting in an estimated OS HR of 0.86 (95% CI: 0.75–0.98; $p = 0.023$); this corresponded to a 1.4-month improvement in median OS from 9.1 months to 10.5 months. Even though this study met its primary end point and was statistically significant, the examination of the CI reveals that ramucirumab confers, at best, an OS HR of 0.75 when compared with control, and just barely excludes the value of one at the upper bound. The study also conveys that statistical significance can be achieved by observing a result that is more modest than the effect size targeted in the statistical design of the study, which is one of the reasons why many researchers believe that studies should be more ambitious in their designs. Nevertheless, the Food and Drug Administration approved ramucirumab for use in combination with docetaxel for this patient population[4]; but regarding clinical relevance, one questions whether the additional risk for adverse events and health care costs outweigh the modest benefit observed for patients in this study.

Statistical significance and clinical significance will usually be in harmony in a well-designed clinical trial with an appropriately chosen target effect and power that minimizes the chances of observing a clinically irrelevant yet statistically significant result.[5] An exception to this is a post hoc analysis of a subset of patients, in which the sample size of the subset may be smaller than that required to detect a clinically meaningful result. Other exceptions arise when nontraditional statistical methods that have been developed to minimize the observed p value of the trial data are used. For example, methods that optimally weight different parts of the survival experience[6,7] may lead to statistically significant results that are not clinically significant.[8] Another example is given by methods that increase the sample size of a trial based on interim results,[9] which can also potentially lead to statistical significance without clinical significance.[10] Although statistically valid, these methods require an especially careful examination of the clinical significance of the results.

Whereas the framework for determination of statistical significance is well established, it is not possible to delineate such a structured framework for the determination of clinical significance, which is inherently a judgmental call made by clinicians and patients. Opinions about the magnitude of effect size required for clinical significance are subjective and may change over time. The American Society of Clinical Oncology's Value Framework or the European Society for Medical Oncology's Magnitude of Clinical Benefit Scale are tools that are updated regularly to help guide the evaluation of the clinical benefit from new cancer therapies.[11,12] Controversy arises when the interpretation of clinical trial results and their impact on patient care is made solely on *p* values; as such, we encourage the use of CIs and consideration of effect sizes, as described above, to provide more clarity on the precision of estimates from a study, and at the same time, acknowledging that overpowered studies or studies with large sample sizes are more easily able to identify trivial but statistically significant differences in outcomes. This is similar to the guidelines put forth by other journals.[13] If the CIs include values that reflect effect sizes that are impactful to patients, then clinical significance could be considered.

References

1. Rubinstein LV, Korn EL, Freidlin B, Hunsberger S, Ivy SP, Smith MA. Design issues of randomized phase II trials and a proposal for randomized phase II screening trials. *J Clin Oncol*. 2005;23:7199-7206.
2. Herbst RS, Baas P, Kim DW, et al. Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial. *Lancet*. 2016;387:1540-1550.
3. Garon EB, Ciuleanu TE, Arrieta O, et al. Ramucirumab plus docetaxel versus placebo plus docetaxel for second-line treatment of stage IV non-small-cell lung cancer after disease progression on platinum-based therapy (REVEL): a multicentre, double-blind, randomised phase 3 trial. *Lancet*. 2014;384:665-673.
4. Larkins E, Scepura B, Blumenthal GM, et al. U.S. Food and Drug Administration approval summary: ramucirumab for the treatment of metastatic non-small cell lung cancer following disease progression on or after platinum-based chemotherapy. *Oncologist*. 2015;20:1320-1325.
5. Cook JA, Fergusson DA, Ford I, et al. There is still a place for significance testing in clinical trials. *Clin Trials*. 2019;16:223-224.
6. Karrison TG. Versatile tests for comparing survival curves based on weighted log-rank statistics. *STATA J*. 2016;16:678-690.
7. Horiguchi M, Cronin AM, Takeuchi M, Uno H. A flexible and coherent test/estimation procedure based on restricted mean survival times for censored time-to-event data in randomized clinical trials. *Stat Med*. 2018;37:2307-2320.
8. Freidlin B, Korn EL. Methods for accommodating nonproportional hazards in clinical trials: ready for the primary analysis? *J Clin Oncol*. 2019;37:3455-3459.
9. Chen YH, Li C, Lan KK. Sample size adjustment based on promising interim results and its application in confirmatory clinical trials. *Clin Trials*. 2015;12:584-595.
10. Freidlin B, Korn EL. Sample size adjustment designs with time-to-event outcomes: a caution. *Clin Trials*. 2017;14:597-604.
11. Schnipper LE, Davidson NE, Wollins DS, et al. Updating the American Society of Clinical Oncology value framework: revisions and reflections in response to comments received. *J Clin Oncol*. 2016;34:2925-2934.
12. Cherny NI, Dafni U, Bogaerts J, et al. ESMO-magnitude of clinical benefit scale version 1.1. *Ann Oncol*. 2017;28:2340-2366.
13. Harrington D, D'Agostino RB Sr, Gatsonis C, et al. New guidelines for statistical reporting in the journal. *N Engl J Med*. 2019;381:285-286.