Interpreting Clinical Trial Outcomes for Optimal Patient Care: A Survey of Clinicians and Trainees

Tanner J. Caverly, MD, MPH

Daniel D. Matlock, MD, MPH

Allan V. Prochazka, MD, MSc

Brian P. Lucas, MD, MS

Rodney A. Hayward, MD

ABSTRACT

Background Evaluation of the clinical importance of outcomes in research studies is an essential element of clinical decision making.

Objective To understand how clinicians and trainees weigh the importance of different types of clinical outcomes in drug trials.

Methods A self-administered paper survey contained 4 scenarios asking participants to rate (1, “no proof” to 10, “good proof”) the extent to which 4 study outcomes provided “proof that the new drug might help people.” Outcomes included (1) a surrogate outcome; (2) a surrogate-enriched composite outcome; (3) stroke mortality; and (4) all-cause mortality. The primary study metrics were mean ratings for each of the 4 outcome types, and the proportion ranking outcome importance of all-cause mortality > stroke mortality > surrogate-enriched composite or surrogate alone.

Results A convenience sample of 549 clinicians and trainees at 2 medical centers completed the survey (response rate: 87% medical students, 80% internal medicine residents, 69% general medicine faculty, and 41% physician experts). The surrogate-enriched composite outcome and stroke mortality were rated the most important evidence for benefit (6.6 and 6.4, respectively), with all-cause mortality and a surrogate outcome being rated significantly lower (5.2 and 4.6, respectively). In addition, 48% of clinicians rated improvement in all-cause mortality as more valuable than an improvement in a surrogate marker. Only 21% rated all-cause mortality as more valuable than a surrogate-enriched composite outcome.

Conclusions These findings raise concerns that clinicians and trainees may not interpret trial evidence in a way that promotes the best care for patients.

Introduction

When evaluating clinical trials for new or established therapies, misjudgments can occur if the clinical importance of the primary outcome is not carefully weighed. Trainees should be taught to ask, “What exactly is the outcome, and how much do my patients care about reducing it?”¹ Careful attention is important because outcomes vary widely in clinical importance.

Clinicians should be aware of nuances in interpreting the value of trial outcomes. First, improvements in surrogate outcomes (eg, cholesterol, blood pressure, or hemoglobin A1c) do not necessarily lead to improvements in outcomes important to patients. Second, outcomes that are a composite of multiple endpoints (composite outcomes) can easily lead to clinical misjudgments about the clinical importance of

an intervention—especially if the composite contains surrogate endpoints (ie, is a surrogate-enriched composite outcome).^{1–3} This occurs because medical interventions often have their largest impact on the least clinically important components of a composite outcome, and a small or nonexistent impact on the most important components.⁴ Finally, disease-specific mortality—in contrast to all-cause mortality—overestimates how much an intervention increases life expectancy (a function of competing risks) and can sometimes obscure that a treatment does net harm.⁵

Despite the importance of appropriately assessing different types of trial outcomes, how clinicians and trainees typically assign importance to these outcomes has not been assessed. The goal of our study was to assess how clinicians and trainees interpret the clinical importance of a treatment, when improvements in 4 different outcomes are offered as evidence of a drug’s potential to help patients: surrogate outcome, surrogate-enriched composite outcome, disease-specific mortality, and all-cause mortality. Since surrogate and surrogate-enriched composite outcomes can be unreliable indicators of clinically important benefits,

DOI: <http://dx.doi.org/10.4300/JGME-D-15-00137.1>

Editor’s Note: The online version of this article contains a table of characteristics of the study sample, the survey tool, and examples of survey scenarios.

we hypothesized that clinicians would rate improvements in these types of outcomes as less important than improvements in mortality.

Methods

Survey Development

The 4 scenarios for this study were part of a larger test evaluating physicians' ability to critically interpret risk information.⁶ The survey tool and information on its development and testing can be found as online supplemental material.

Study Design

We targeted convenience samples of clinicians and trainees at 2 institutions, a large academic medical institution and a university-affiliated community hospital, with most of the sample coming from the first institution. Research staff distributed the self-administered survey to trainees in attendance at core educational conferences, and mailed surveys to faculty in the division of general internal medicine.

Each participant taking the survey received the same 4 scenarios describing the benefits derived from a new drug. Scenarios were always in the same order and were similar except for the type of outcome reported (provided as online supplemental material). Participants were asked to rate the extent to which each scenario provided proof that the new drug "might help people." We chose nonspecific phrasing intentionally. Although we wished to make it clear that we were referring to clinical significance, not statistical significance, we felt it was critical that survey respondents determine for themselves what is meant by a medication "helping people." Answers used a 10-point response scale from 1 to 10 with 2 anchors (1, "no proof" and 10, "good proof"). Each scenario was separated by a number of distracter questions to limit direct comparisons, in order to better reflect what might happen when a practicing clinician reads individual articles in the medical literature.

Participants answered basic demographic and clinical practice questions, and they completed a previously developed test of statistical numeracy with moderate validity evidence (the 4-item Berlin Numeracy Test).⁷ Statistical numeracy is defined as "the ability to accurately interpret and act on information about risk."

Setting

The surveys were distributed at both participating institutions. Evidence-based medicine at these 2

What was known and gap

Appropriately interpreting evidence is important to ascertain the best care for patients, but it is not known how clinicians ascribe importance to different types of clinical outcomes.

What is new

A sample of 549 clinicians, including trainees, rated whether 4 study outcomes provided "proof a new drug might help people."

Limitations

Convenience sample limits generalizability; participants had cues about the presence of outcomes with higher clinical relevance.

Bottom line

Clinicians may not interpret clinical trial outcomes in ways that promote good care for patients.

institutions was typically taught during several lectures and small group sessions for medical students, and as part of monthly journal clubs with 1 or 2 clinical rotations with formal lectures per year for internal medicine residents.

Participants

The sample included third-year and fourth-year medical students, internal medicine residents, and faculty in the division of general internal medicine at the first institution, as well as a group of internal medicine interns at the second institution. In addition, a national group of clinician-researchers with evidence-based medicine expertise, who review health efficacy and safety claims for HealthNewsReview.org,⁸ took an online version of the survey.

The Colorado Institutional Review Board approved this anonymous survey as exempt human research with waiver of informed consent.

Analysis

We used descriptive statistics to summarize how the 4 scenarios were rated overall and to describe how individual participants rated 1 scenario relative to the other 3. Five paired *t* tests were used to compare how mean ratings on the 1- to 10-point scale differed between the 4 different types of outcomes ($P < .01$ was considered statistically significant after Bonferroni correction). We calculated a standardized mean difference—dividing mean differences in how scenarios were rated by the pooled SD. Based on a normative hierarchy of clinical importance,¹ we created a score for how each clinician ranked the value of one outcome relative

TABLE 1
The Relative Value Clinicians Place on Different Trial Outcomes^{a,b}

Category	Question	Mean Rating	95% CI
Composite	In a large randomized trial, people in the new drug group experienced decreased rates of the combined primary endpoint (nonfatal stroke, death, or elevated levels of a risk factor into the high-risk category for stroke).	6.6	6.4–6.8
Disease-specific mortality	In a large randomized controlled trial, fewer people died from stroke in the new drug group than in the placebo group.	6.4	6.2–6.5
All-cause mortality	In a large randomized trial, fewer people died for any reason in the new drug group than in the placebo group.	5.2	5.0–5.4
Surrogate	A large randomized trial showed that a new drug lowers serum levels of a risk factor known to be associated with an increased risk of death from stroke.	4.6	4.4–4.7

^a Participants were told that all of the results were statistically significant from randomized trials with excellent validity and generalizability, and they were asked to “rate the extent to which this provides proof that the new drug *might* help people” (1, “no proof” to 10, “good proof”).

^b The comparison between “composite” and “disease-specific mortality” had $P = .02$ and was considered nonsignificant ($P < .01$ after Bonferroni correction for multiple comparisons). Standardized mean differences for 3 of the comparisons indicated a “moderate” effect size (stroke versus surrogate, $SMD = 0.68$; all-cause versus composite, $SMD = 0.51$; and all-cause versus stroke, $SMD = 0.47$). The effect size was small between the all-cause versus surrogate scenarios ($SMD = 0.22$).

to another, assigning 1 point to each “correct” ranking: (1) stroke mortality > surrogate; (2) all-cause mortality > surrogate; (3) stroke mortality > composite; (4) all-cause mortality > composite; and (5) all-cause mortality > stroke mortality. “Incorrect” rankings and missing responses received 0 points. This yielded a 0 to 5 summary score for each clinician. Finally, ordinal regression—with the score from 0 to 5 points as the dependent variable—was used to identify how the scores varied across professional groups and numeracy scores, controlling for sex and number of additional degrees. Adjusted probabilities of scoring 0, 1, 2, 3, 4, or 5 points across professional groups and numeracy scores were then derived from the ordinal regression model. All analyses were completed using STATA version 13.1 (StataCorp LP, College Station, TX).

Results

We received 549 completed surveys for analysis. Participants included 258 women (47%), 273 medical students (50%), 148 internal medicine residents (27%), 120 general medicine faculty (22%), and 7 physician experts (1%). The respective response rates were 87% (273 of 313) medical students, 80% (148 of 185) internal medicine residents, 69% (120 of 175) academic general internists, and 41% (7 of 17) physician experts. Additional characteristics of the survey participants are provided as online supplemental material.

Distribution of Overall Mean Responses

TABLE 1 shows mean responses for the 4 scenarios on the 1 to 10 scale (1, no proof of benefit, to 10, good proof of benefit). On average, participants rated improvement in the surrogate-enriched composite outcome (containing a surrogate endpoint as well as mortality component endpoints) as the best proof that a drug might “help people” (mean = 6.6). Although improvement in the surrogate outcome was rated lower (mean = 4.6), all-cause mortality was rated only slightly higher (mean = 5.2). All-cause mortality was rated lower than both the surrogate-enriched composite outcome and stroke mortality (mean = 6.4). All pairwise comparisons between the mean ratings were statistically significant at $P < .01$ on paired t tests, except for the difference between the composite outcome and stroke mortality ratings ($P = .02$).

Standardized mean differences (SMDs) were calculated to help interpret the importance of differences in means across the scenarios. Typically, SMDs around 0.2 represent small changes, those around 0.5 represent moderate changes, and those around 0.8 represent large changes.² The SMDs were in the range of 0.5 for 3 of the comparisons (stroke versus surrogate, $SMD = 0.68$; all-cause versus composite, $SMD = 0.51$; and all-cause versus stroke, $SMD = 0.47$), indicating “moderate” effect sizes across these pairs (TABLE 1).² The effect size was small between the all-cause versus surrogate scenarios ($SMD = 0.22$).

TABLE 2
Proportion of Clinicians Ranking Comparisons “Correctly”^a

Pairwise Comparison (n) ^b	Proportion, %
Stroke mortality > surrogate (n = 547)	70
All-cause mortality > surrogate (n = 545)	48
Stroke mortality > composite (n = 548)	31
All-cause mortality > composite (n = 544)	21
All-cause mortality > stroke mortality (n = 545)	19

^a “Correctly” defined by the following pairwise rankings: stroke mortality > surrogate; all-cause mortality > surrogate; stroke mortality > composite; all-cause mortality > composite; and all-cause mortality > stroke mortality.

^b Numbers vary due to item nonresponse.

Pairwise Comparisons

Since each participant answered all 4 scenarios, we were able to assess how individuals responded on one scenario relative to their response on another scenario. TABLE 2 presents the proportion of clinicians and trainees answering each pairwise comparison “correctly” based on a normative hierarchy of clinical importance: (1) stroke mortality > surrogate; (2) all-cause mortality > surrogate; (3) stroke mortality > composite; (4) all-cause mortality > composite; and (5) all-cause mortality > stroke mortality.¹ Only about half (48%, 263 of 545) rated improvement in all-cause mortality as better proof that a new drug “might help people” than an improvement in a surrogate marker, only 21% (112 of 544) rated improvement in all-cause mortality more highly than improvement in a surrogate-enriched composite outcome, and only 19% (105 of 545) rated improvement in all-cause mortality more highly than an improvement in stroke-related mortality. Overall, 29% (156 of 544) of the participants correctly ordered 3 to 5 outcome comparisons, while 4% (20 of 544) ordered all 5 outcome pairs “correctly.”

Effect of Training and Statistical Numeracy on Interpretation

The FIGURE presents differences in “correct” ratings across subgroups. After adjusting for sex, number of additional degrees, and statistical numeracy in the multivariable ordinal model, there were no significant differences in the number of “correct” ratings given by students, residents, and faculty. However, the small

group of 7 physician experts was significantly more likely to rank outcome comparisons “correctly” than the other groups in the adjusted ordered logistic regression model ($P = .005$; FIGURE). Those answering 4 of 4 Berlin Numeracy Test questions correctly were also significantly more likely to correctly order outcome pairs ($P = .04$).

Discussion

In this vignette-based study of a convenience sample of trainees and practicing academic physicians to assess the clinical importance placed on 4 different types of trial outcomes, we found that many overrated the importance of surrogate outcomes and even more overrated surrogate-enriched composite outcomes. At the same time, improvements in all-cause mortality were underrated by most participants. The small group of physician experts and those with higher levels of statistical numeracy were somewhat more likely to correctly order the 4 outcomes types. One would expect different results if participants were appropriately sensitive to the clinical importance of different trial outcomes.

These results raise concerns that clinicians and trainees, when weighing evidence for clinical decisions, may not adequately consider the clinical importance of the type of outcome being reported. This could lead to the misallocation of health care resources toward less valuable interventions that may not substantially improve health. In the worst case scenario, reliance on improving surrogate outcomes could lead to pursuing interventions that cause net harm. The recent case of rosiglitazone provides a cautionary tale.^{4,9}

Our study should be interpreted with several limitations in mind. We used a convenience sample at 2 institutions, and the results may not generalize to other populations. Response rates were generally good, but the sample included just a small number of evidence-based medicine physician experts. Although the 4 scenarios analyzed here were separated by distracter questions, direct comparisons between scenarios were possible. This could potentially lead to more deliberate responses than would occur in routine clinical practice, where physicians form judgments about the importance of a study outcome in the absence of cues that more clinically important outcomes were not demonstrated.

Disease-specific mortality is usually more sensitive (offers greater statistical power) than all-cause mortality for observing treatment effects.¹⁰ This may have resulted in some respondents rating stroke mortality more highly than all-cause mortality.

FIGURE

Combined Probability of Ranking 3, 4, or 5 Comparisons Correctly^a by Clinical Background and Statistical Numeracy

^a “Correctly” defined by the following pairwise rankings: stroke mortality > surrogate; all-cause mortality > surrogate; stroke mortality > composite; all-cause mortality > composite; and all-cause mortality > stroke mortality. The combined probability of answering 3, 4, or 5 pairwise comparisons “correctly” is presented (ie, “majority correct”).

^b $P < .05$ in the ordinal logistic regression model.

However, we did ask respondents about which outcome is more important in determining that treatment “might help people.” For that determination, all-cause mortality is the more important outcome. Similarly, our wording was meant to steer respondents away from ratings based on evidence of biological plausibility. It remains possible that some respondents rated all-cause mortality lower because of biological implausibility, or the erroneous belief that biological implausibility indicates a lack of clinical importance. Finally, it is possible that some participants believe that dying from a stroke is more problematic than dying from most other causes, and thus rated improving stroke mortality as more important than improving all-cause mortality. Cognitive “think-aloud” interviews with test-takers would be needed to further explore why participants rated the different questions as they did.

These findings suggest that clinicians and trainees may uncritically accept as valuable drugs that

improve only surrogate outcomes, but do not provide the best care for patients. This problem could certainly be addressed through medical education. The Alliance for Academic Internal Medicine–American College of Physicians High-Value Curriculum,¹¹ and other recent educational initiatives, are well positioned to include teaching about the pitfalls of surrogate and composite outcomes. Our findings also add support to efforts that promote more transparent reporting of clinical information to clinicians. For example, labels for drugs approved on the basis of surrogate outcomes should report the lack of evidence for improving clinically important outcomes.

Conclusion

The clinicians and trainees had difficulty appropriately rating outcomes that indicate unambiguous clinical importance more highly than outcomes of

uncertain clinical importance. General medicine faculty were no more likely to rate outcome types appropriately than residents or medical students.

References

1. Woloshin W, Schwartz LM, Welch HG. *Know Your Chances: Understanding Health Statistics*. 1st ed. Berkeley: University of California Press; 2008.
2. Guyatt G. *JAMA's Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice*. New York, NY: McGraw-Hill Medical; 2008.
3. Cordoba G, Schwartz L, Woloshin S, Bae H, Gotzsche PC. Definition, reporting, and interpretation of composite outcomes in clinical trials: systematic review. *BMJ*. 2010;341:c3920.
4. Ferreira-Gonzalez I, Permyer-Miralda G, Domingo-Salvany A, Busse JW, Heels-Ansdell D, Montori VM, et al. Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials. *BMJ*. 2007;334(7597):786.
5. Black WC, Haggstrom DA, Welch HG. All-cause mortality in randomized trials of cancer screening. *J Natl Cancer Inst*. 2002;94(3):167–173.
6. Caverly TJ, Prochazka AV, Combs BP, Lucas BP, Mueller SR, Kutner JS, et al. Doctors and numbers: an assessment of the critical risk interpretation test. *Med Decis Making*. 2015;35(4):512–524.
7. Cokely ET, Galesic M, Schulz E, Ghazal S, Garcia-Retamero R. Measuring risk literacy: The Berlin Numeracy Test. *Judgm Decis Making*. 2012;7(1):25–47.
8. Schwitzer G. A guide to reading health care news stories. *JAMA Intern Med*. 2014;174(7):1183–1186.
9. Nissen SE. The rise and fall of rosiglitazone. *Eur Heart J*. 2010;31(7):773–776.
10. Yusuf S, Negassa A. Choice of clinical outcomes in randomized trials of heart failure therapies: disease-specific or overall outcomes? *Am Heart J*. 2002;143(1):22–28.
11. Smith CD. Teaching high-value, cost-conscious care to residents: The Alliance for Academic Internal Medicine–American College of Physicians Curriculum. *Ann Intern Med*. 2012;157(4):284–286.

Tanner J. Caverly, MD, MPH, is Research Scientist, Veterans Affairs Center for Clinical Management Research, and Clinical Lecturer, Department of Internal Medicine and Department of Learning Health Sciences, University of Michigan Medical School; **Daniel D. Matlock, MD, MPH**, is Associate Professor of Medicine, University of Colorado School of Medicine; **Allan V. Prochazka, MD, MSc**, is Professor of Medicine, University of Colorado School of Medicine and Denver Veterans Affairs Health System; **Brian P. Lucas, MD, MS**, is Associate Professor of Medicine, Veterans Affairs Medical Center, and Geisel School of Medicine, Dartmouth College; and **Rodney A. Hayward, MD**, is Senior Research Scientist, Veterans Affairs Center for Clinical Management Research, and Professor of Medicine, Department of Internal Medicine, University of Michigan Medical School.

Funding: This study used the Methods Core of the Michigan Center for Diabetes Translational Research (NIDDK P30DK092926) and also HX 13-001 (VA HSR&D Center of Innovation).

Conflict of interest: The authors declare they have no competing interests.

This manuscript was presented at the Society for General Internal Medicine Annual Conference, San Diego, California, April 22, 2014.

Corresponding author: Tanner J. Caverly, MD, MPH, Veterans Affairs Medical Center, I11D, 2215 Fuller Road, Ann Arbor, MI 48105, 734.222.8958, tcaverly@med.umich.edu

Received March 23, 2015; revision received August 2, 2015; accepted September 14, 2015.