

Table of contents:

1. The Intention-to-Treat Principle
2. Noninferiority Trials
3. Sample Size Calculation for a Hypothesis Test
4. Interpretation of Clinical Trials That Stopped Early
5. Cluster Randomized Trials
6. Case-Control Studies
7. Decision Curve Analysis
8. Gatekeeping Strategies for Avoiding False-Positive Results in Clinical Trials With Many Comparisons
9. Multiple Comparison Procedures
10. Pragmatic Trials
11. Equipoise in Research
12. The Propensity Score
13. Dose-Finding Trials
14. Odds Ratios—Current Best Practice and Use
15. Evaluating Discrimination of Risk Prediction Models
16. Time-to-Event Analysis
17. The Stepped-Wedge Clinical Trial
18. Mendelian Randomization
19. Bayesian Analysis: Using Prior Information to Interpret the Results of Clinical Trials

JAMA Guide to Statistics and Methods

The Intention-to-Treat Principle

How to Assess the True Effect of Choosing a Medical Treatment

Michelle A. Detry, PhD; Roger J. Lewis, MD, PhD

The intention-to-treat (ITT) principle is a cornerstone in the interpretation of randomized clinical trials (RCTs) conducted with the goal of influencing the selection of medical therapy for well-defined groups of patients. The ITT principle defines both the study



Related article 36

population included in the primary efficacy analysis and how the outcomes are analyzed. Under ITT, study participants are analyzed as members of the treatment group to which they were randomized regardless of their adherence to, or whether they received, the intended treatment.¹⁻³ For example, in a trial in which patients are randomized to receive either treatment A or treatment B, a patient may be randomized to receive treatment A but erroneously receive treatment B, or never receive any treatment, or not adhere to treatment A. In all of these situations, the patient would be included in group A when comparing treatment outcomes using an ITT analysis. Eliminating study participants who were randomized but not treated or moving participants between treatment groups according to the treatment they received would violate the ITT principle.

In this issue of *JAMA*, Robertson et al conducted an RCT using a factorial design to compare transfusion thresholds of 10 and 7 g/dL and administration of erythropoietin vs placebo in 895 patients with anemia and traumatic brain injury.⁴ The primary outcome was the 6-month Glasgow Outcome Scale (GOS), dichotomized so a good or moderate score indicated success. The trial was conducted with high fidelity to the protocol so only a few patients did not receive the intended treatment strategy. Two patients randomized to the 7-g/dL study group were managed according to the 10-g/dL threshold and an additional 2 patients randomized to the 7-g/dL study group received one transfusion not according to protocol. The authors implemented the ITT principle and the outcomes for these 4 patients were included in the 7-g/dL group.

Use of the Method

Why Is ITT Analysis Used?

The effectiveness of a therapy is not simply determined by its pure biological effect but is also influenced by the physician's ability to administer, or the patient's ability to adhere to, the intended treatment. The true effect of selecting a treatment is a combination of biological effects, variations in compliance or adherence, and other patient characteristics that influence efficacy. Only by retaining all patients intended to receive a given treatment in their original treatment group can researchers and clinicians obtain an unbiased estimate of the effect of selecting one treatment over another.

Treatment adherence often depends on many patient and clinician factors that may not be anticipated or are impossible to measure and that influence response to treatment. For example, in the study by Robertson et al, some patients randomized to the higher transfusion threshold may not have received the intended therapeutic strategy due to adverse events associated with transfusion, fluid overload, or unwillingness of clinicians to adhere to the strategy for other reasons. These patients are likely to be fundamentally different from those who were actually treated using the 10-g/dL strategy. The characteristics that differ between patients who received the intended therapy and those who did not could easily influence whether a successful GOS score is achieved. If the ITT principle was not followed and patients were removed from their randomized group and either ignored or assigned to the other treatment group, the results of the analysis would be biased and no longer represent the effect of choosing one therapy over the other.

It is common to see alternative analyses proposed, eg, per-protocol or modified intent-to-treat (MITT) analyses.⁵ A per-protocol analysis includes only study participants who completed the trial without any major deviations from the study protocol; this usually requires that they successfully receive and complete their assigned treatment(s), complete their study visits, and provide primary outcome data. The requirements to be included in the per-protocol analysis vary from study to study. While the definition of an MITT analysis also varies from study to study, the MITT approach deviates from the ITT approach by eliminating patients or reassigning patients to a study group other than the group to which they were randomized. Neither of these approaches satisfies the ITT principle and may lead to clinically misleading results. It has been observed that studies using MITT analysis are more likely to be positive than those following a strict ITT approach.⁵ A comparison of results from ITT and per-protocol or MITT analyses may provide some indication of the potential effect of nonadherence on overall treatment effectiveness.

Noninferiority trials, which are designed to demonstrate that an experimental treatment is no worse than an established one, require special considerations with regard to the ITT principle.⁶⁻⁸ Consider a noninferiority trial of 2 treatments—treatment A is a biologically ineffective experimental therapy and treatment B is a biologically effective standard therapy—with the goal to demonstrate that treatment A is noninferior to B. Patients may be randomized to receive treatment B, not adhere to the treatment, and fail treatment due to their nonadherence. If this happens frequently, treatment B will appear less efficacious. Thus, the intervention in group A may incorrectly appear noninferior to the intervention in group B, simply as a result of nonadherence rather than because of similar biological efficacy. In this case, the ITT

analysis is somewhat misleading because the noninferiority is a result of poor adherence. In a noninferiority trial, both ITT and per-protocol analyses should be conducted and reported. If the per-protocol results are similar to the ITT results, the claim of noninferiority is substantially strengthened.⁶⁻⁸

What Are the Limitations of ITT Analysis?

A characteristic of the ITT principle is that poor treatment adherence may result in lower estimates of treatment efficacy and a loss of study power. However, these estimates are clinically relevant because real-world effectiveness is limited by the ability of patients and clinicians to adhere to a treatment.

Because all patients must be analyzed under the ITT principle, it is essential that all patients be followed up and their primary outcomes determined. Patients who discontinue study treatments are often more likely to be lost to follow-up. Following the ITT principle will not eliminate bias associated with missing outcome data; steps must always be taken to keep missing data to a minimum and, when missing data are unavoidable, to use minimally biasing methods for adjusting for missing data (eg, multiple imputation).

Why Did the Authors Use ITT Analysis in This Particular Study?

Robertson et al⁴ used an ITT analysis because it allowed the effectiveness of their therapeutic strategies to be evaluated without bias due to differences in adherence. Failure to follow the ITT principle could have led to greater scrutiny of the trial results, especially if adherence to the intended treatments had been poorer.

Caveats to Consider When Looking at Results Based on ITT Analysis

Although the ITT principle is important for estimating the efficacy of treatments, it should not be applied in the same way in assessing the safety (eg, medication adverse effects) of interventions. For example, it would not make sense to attribute an apparent adverse effect to an intended treatment when, in fact, the patient was never exposed to the experimental drug. For this reason, safety analyses are generally conducted according to the treatment actually received, even though this may not accurately estimate—and may well overestimate—the burden of adverse effects likely to be seen in clinical practice.

While determining the effect of choosing one treatment over another, or over no treatment at all, is a key goal of trials conducted late in the process of drug and device development, the goals of trials conducted earlier in development are generally focused on narrower questions such as biological efficacy and dose selection. In these cases, MITT and per-protocol analysis strategies have a greater role in guiding the design and conduct of subsequent clinical trials. For example, it would be unfortunate to falsely conclude, based on the ITT analysis of a phase 2 clinical trial, that a novel pharmaceutical agent is not effective when, in fact, the lack of efficacy stems from too high a dose and patients' inability to be adherent because of intolerable adverse effects. In that case, a lower dose may yield clinically important efficacy and a tolerable adverse effect profile. A per-protocol analysis may be helpful in such a case, allowing the detection of the beneficial effect in patients able to tolerate the new therapy.

ARTICLE INFORMATION

Author Affiliations: Berry Consultants LLC, Austin, Texas (Detry, Lewis); Department of Emergency Medicine, Harbor-UCLA Medical Center; Los Angeles Biomedical Research Institute; and David Geffen School of Medicine at UCLA, Torrance, California (Lewis).

Corresponding Author: Roger J. Lewis, MD, PhD, Department of Emergency Medicine, Harbor-UCLA Medical Center, Bldg D9, 1000 W Carson St, Torrance, CA 90509 (roger@emedharbor.edu).

Conflict of Interest Disclosures: The authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

REFERENCES

1. Cook T, DeMets DL. *Introduction to Statistical Methods for Clinical Trials*. Boca Raton, FL:

Chapman & Hall/CRC; Taylor & Francis Group; 2008:chp 11.

2. Schulz KF, Altman DG, Moher D; CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. *Ann Intern Med*. 2010;152(11):726-732.

3. Food and Drug Administration. Guidance for industry e9 statistical principles for clinical trials. <http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm073137.pdf>. September 1998. Accessed May 11, 2014.

4. Robertson CS, Hannay HJ, Yamal J-M, et al; and the Epo Severe TBI Trial Investigators. Effect of erythropoietin and transfusion threshold on neurological recovery after traumatic brain injury: a randomized clinical trial. *JAMA*. doi:10.1001/jama.2014.6490.

5. Montedori A, Bonacini MI, Casazza G, et al. Modified versus standard intention-to-treat reporting: are there differences in methodological quality, sponsorship, and findings in randomized trials? a cross-sectional study. *Trials*. 2011;12:58.

6. Piaggio G, Elbourne DR, Pocock SJ, Evans SJ, Altman DG; CONSORT Group. Reporting of noninferiority and equivalence randomized trials: extension of the CONSORT 2010 statement. *JAMA*. 2012;308(24):2594-2604.

7. Le Henanff A, Giraudeau B, Baron G, Ravaud P. Quality of reporting of noninferiority and equivalence randomized trials. *JAMA*. 2006;295(10):1147-1151.

8. Mulla SM, Scott IA, Jackevicius CA, et al. How to use a noninferiority trial: Users' Guides to the Medical Literature. *JAMA*. 2012;308:2605-2611.

Noninferiority Trials

Is a New Treatment Almost as Effective as Another?

Amy H. Kaji, MD, PhD; Roger J. Lewis, MD, PhD

Sometimes the goal of comparing a new treatment with a standard treatment is not to find an approach that is more effective but to find a therapy that has other advantages, such as lower cost, fewer



Related article page 2340

adverse effects, or greater convenience with at least similar efficacy to the standard treatment. With other advantages, a treatment that is almost as effective as a standard treatment might be preferred in practice or for some patients. The purpose of a noninferiority trial is to rigorously evaluate a new treatment against an accepted and effective treatment with the goal of demonstrating that it is at least almost as good (ie, not inferior).

In this issue of *JAMA*, Salminen et al describe the results of a multicenter noninferiority trial of 530 adults with computed tomography–confirmed acute appendicitis who were randomized either to early appendectomy (the standard treatment) or to antibiotic therapy alone (a potentially less burdensome experimental treatment).¹

Use of the Method

Why Are Noninferiority Trials Conducted?

In a traditional clinical trial, a new treatment is compared with a standard treatment or placebo with the goal of demonstrating that the new treatment has greater efficacy. The null hypothesis for such a trial is that the 2 treatments have the same effect. Rejection of this hypothesis, implying that the effects are different, is signaled by a statistically significant *P* value or, alternatively, by a 2-tailed confidence interval that excludes no effect. While the new treatment could be either superior or inferior, the typical trial aims to demonstrate superiority of the new treatment and is known as a superiority trial. Since a superiority trial is capable of identifying both harmful and beneficial effects of a new therapy vs a control (ie, a current therapy), a 2-tailed 95% CI can be used to indicate the upper and lower limits of the difference in treatment effect that are consistent with the observed data. The null hypothesis is rejected, indicating that the new therapy differs from the control, if the confidence interval does not include the result that indicates absence of effect (eg, a risk ratio of 1 or a risk difference of 0).² This is equivalent to a statistically significant *P* value.

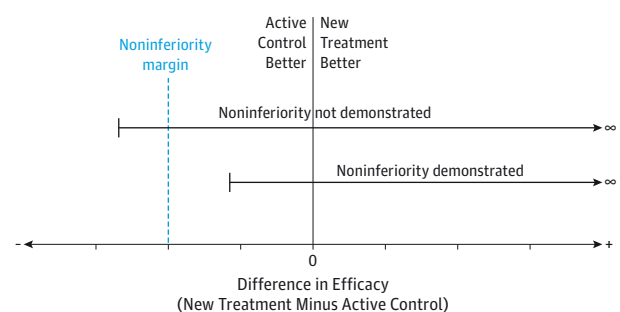
Although superiority or inferiority of a new treatment can be demonstrated by a superiority trial, it would generally be incorrect to conclude that the absence of a significant difference in a superiority trial demonstrates that the therapies have similar effects; absence of evidence of a difference is not reliable evidence that there is no difference. An active-controlled noninferiority trial is needed to determine whether a new intervention, which offers other advantages such as decreased toxicity or cost, does not have lesser efficacy than an established treatment.³⁻⁶ Noninferiority trials use

known effective treatments as controls because there is little to be gained by demonstrating that a new therapy is not inferior to a sham or placebo treatment.

The objective of a noninferiority trial is to demonstrate that the intervention being evaluated achieves the efficacy of the established therapy within a predetermined acceptable noninferiority margin (Figure). The magnitude of this margin depends on what would be a clinically important difference, the expected event rates, and, possibly, regulatory requirements. Other determinants of the noninferiority margin include the known effect of the standard treatment vs placebo; the severity of the disease; toxicity, inconvenience, or cost of the standard treatment; and the primary end point. A smaller noninferiority margin is likely appropriate if the disease under investigation is severe or if the primary end point is death.³⁻⁶

The sample size required to reliably demonstrate noninferiority depends on both the choice of the noninferiority margin and the assumed true relative effects of the compared treatments.³⁻⁶ An active-controlled noninferiority trial often requires a larger sample size than a superiority trial because the noninferiority margins used in noninferiority studies are generally smaller than the differences sought in superiority trials. Just as important is the assumed effect of the experimental treatment relative to the active-control treatment. The assumed effect may be that the experimental treatment is worse than the control but by a smaller amount than the noninferiority margin, that the 2 treatments are equivalent, or even that

Figure. Two Different Possible Results of a Noninferiority Trial, Summarized by 1-Tailed Confidence Intervals for the Relative Efficacy of the New and Active-Control Treatments



In the top example, the lower limit of the confidence interval lies to the left of the noninferiority margin, demonstrating that the results are consistent with greater inferiority (worse efficacy) than allowed by the noninferiority margin. Thus, the new treatment may be inferior and noninferiority is not demonstrated. In the lower example, the lower limit of the confidence interval lies to the right of the noninferiority margin, demonstrating noninferiority of the new treatment relative to the active-control treatment. The overall result of the trial is defined by the lower limit of the 1-sided confidence interval rather than by the point estimate for the treatment effect, so point estimates are not shown.

the experimental treatment is more effective. These 3 options will result in larger, intermediate, and smaller required sample sizes, respectively, to achieve the same trial power—the chance of demonstrating noninferiority—because they assume progressively better efficacy of the experimental treatment.

Because a noninferiority trial only aims to demonstrate noninferiority and does not aim to distinguish noninferiority from superiority, it is analyzed using a 1-sided confidence interval (Figure) or hypothesis test. Typically, a 1-sided 95% or 97.5% CI (–L to ∞; negative values represent inferiority of the experimental treatment) is constructed for the difference between the 2 treatments, and the lower limit, –L, is compared with the noninferiority margin. Noninferiority is demonstrated if the lower confidence limit lies above or to the right of the noninferiority margin.³⁻⁶

What Are the Limitations of Noninferiority Trials?

A negative noninferiority trial does not in general demonstrate inferiority of the experimental treatment, just as a negative superiority trial does not demonstrate equivalence of 2 treatments.

A noninferiority trial is similar to an equivalence trial in that the objective of both is to demonstrate that the intervention matches the action of the established therapy within a prespecified margin. However, the objective of a noninferiority trial is only to demonstrate that the experimental treatment is not substantially worse than the standard treatment, whereas that of an equivalence trial is to demonstrate that the experimental treatment is neither worse than nor better than the standard treatment.³

Why Was a Noninferiority Trial Conducted in This Case?

Ever since McBurney demonstrated reduced morbidity from pelvic infections with appendectomy, the standard treatment for acute appendicitis has been surgery, which requires general anesthesia, incurs increased cost, and is associated with postoperative complications, such as wound infections and adhesions. Thus, a less invasive approach with similar efficacy might be preferred by many patients and physicians. Three randomized trials summarized in a recent Cochrane analysis demonstrated equipoise as to whether appendicitis can successfully be treated with antibiotics alone rather than surgery.⁷ Because appendectomy is viewed as the standard treatment, it was considered the active control with which the less invasive experimental antibiotic treatment was to be compared.

To design the clinical trial, Salminen et al assumed a surgical treatment success rate of 99% and prespecified a noninferiority margin of –24% based on clinical considerations. This is equivalent to saying that if the rate of treatment success with antibiotics alone could be shown to be no worse than 24% worse than the rate with surgery, then the antibiotic-only strategy would be clinically noninferior. As this study demonstrates, the selection of the noninferiority margin is often subjective rather than based on specific criteria.

How Should the Results Be Interpreted?

The results demonstrated that all but 1 of 273 patients randomized to the surgery group underwent successful appendectomy, resulting in a treatment efficacy of 99.6%. In the antibiotic treatment group, 186 of 256 patients available for follow-up had treatment successes, for a success rate of 72.7%; 70 of the 256 patients underwent surgical intervention within 1 year of initial presentation. Thus, the point estimate for the difference in success rate with the antibiotic-only strategy was –27.0% and the associated 1-tailed 95% CI would range from –31.6% to infinity. Because that interval includes efficacy values worse than the noninferiority margin of –24%, noninferiority cannot be demonstrated.

Caveats to Consider When Looking at a Noninferiority Trial

Noninferiority active-controlled trials often require a larger sample size than placebo-controlled trials, in part because the chosen noninferiority margins are often small. The required sample size for a noninferiority trial is highly dependent on the noninferiority margin and the assumed effect of the new treatment; this assumed effect must be clearly stated and realistic.

The primary analysis for a superiority trial should be based on the intention-to-treat (ITT) principle because it is generally conservative in the setting of imperfect adherence to treatment. However, analyzing a noninferiority trial by ITT could make an inferior treatment appear to be noninferior if poor patient adherence resulted in both treatments being similarly ineffective. Thus, when analyzing a noninferiority trial, both ITT and per-protocol analyses should be conducted. The results are most meaningful when both approaches demonstrate noninferiority.

A noninferiority trial does not distinguish between a new treatment that is noninferior and one that is truly superior and cannot demonstrate equivalence.

ARTICLE INFORMATION

Author Affiliations: Department of Emergency Medicine, Harbor-UCLA Medical Center, Torrance, California (Kaji, Lewis); David Geffen School of Medicine at UCLA, Torrance, California (Kaji, Lewis); Los Angeles Biomedical Research Institute, Los Angeles, California (Kaji, Lewis); Berry Consultants, LLC (Lewis).

Corresponding Author: Roger J. Lewis, MD, PhD, Department of Emergency Medicine, Harbor-UCLA Medical Center, 1000 W Carson St, Box 21, Torrance, CA 90509 (roger@emedharbor.edu).

Section Editors: Roger J. Lewis, MD, PhD, Department of Emergency Medicine, Harbor-UCLA Medical Center and David Geffen School of Medicine at UCLA; and Edward H. Livingston, MD, Deputy Editor, JAMA.

Conflict of Interest Disclosures: The authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

REFERENCES

- Salminen P, Paajanen H, Rautio T, et al. Antibiotic therapy vs appendectomy for treatment of uncomplicated acute appendicitis: the APPAC randomized clinical trial. *JAMA*. doi: 10.1001/jama.2015.6154.
- Young KD, Lewis RJ. What is confidence? I: the use and interpretation of confidence intervals. *Ann Emerg Med*. 1997;30(3):307-310.
- Kaji AH, Lewis RJ. Are we looking for superiority, equivalence, or noninferiority? *Ann Emerg Med*. 2010;55(5):408-411.

4. Mulla SM, Scott IA, Jackevicius CA, et al. How to use a non-inferiority trial: Users' Guides to the Medical Literature. *JAMA*. 2012;308:2605-2611.

5. Tamayo-Sarver JH, Albert JM, Tamayo-Sarver M, Cydulka RK. Advanced statistics: how to determine whether your intervention is different, at least as effective as, or equivalent. *Acad Emerg Med*. 2005;12(6):536-542.

6. Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJW; CONSORT Group. Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. *JAMA*. 2006;295(10):1152-1160.

7. Wilms IM, de Hoog DE, de Visser DC, Janzing HM. Appendectomy vs antibiotic treatment for acute appendicitis. *Cochrane Database Syst Rev*. 2011;(11):CD008359.

JAMA Guide to Statistics and Methods

Sample Size Calculation for a Hypothesis Test

Lynne Stokes, PhD

In this issue of *JAMA*, Koegelenberg et al¹ report the results of a randomized clinical trial (RCT) that investigated whether treatment with a nicotine patch in addition to varenicline produced



Related article page 155

higher rates of smoking abstinence than varenicline alone. The primary results were positive; that is, patients receiving the combination therapy were more likely to achieve continuous abstinence at 12 weeks than patients receiving varenicline alone. The absolute difference in the abstinence rate was estimated to be approximately 14%, which was statistically significant at level $\alpha = .05$.

These findings differed from the results reported in 2 previous studies^{2,3} of the same question, which detected no difference in treatments. What explains this difference? One explanation offered by the authors is that the previous studies "...may have been inadequately powered," which means the sample size in those studies may have been too small to identify a difference between the treatments tested.

Use of the Method

Why Is Power Analysis Used?

The sample size in a research investigation should be large enough that differences occurring by chance are rare but should not be larger than necessary, to avoid waste of resources and to prevent exposure of research participants to risk associated with the interventions. With any study, but especially if the study sample size is very small, any difference in observed rates can happen by chance and thus cannot be considered statistically significant.

In developing the methods for a study, investigators conduct a power analysis to calculate sample size. The power of a hypothesis test is the probability of obtaining a statistically significant result when there is a true difference in treatments. For example, suppose, as Koegelenberg et al¹ did, that the smoking abstinence rate were 45% for varenicline alone and 14% larger, or 59%, for the combination regimen. Power is the probability that, under these conditions, the trial would detect a difference in rates large enough to be statistically significant at a certain level α (ie, α is the probability of a type I error, which occurs by rejecting a null hypothesis that is actually true).

Power can also be thought of as the probability of the complement of a type II error. If we accept a 20% type II error for a difference in rates of size d , we are saying that there is a 20% chance that we do not detect the difference between groups when the difference in their rates is d . The complement of this, $0.8 = 1 - 0.2$, or the statistical power, means that when a difference of d exists, there is an 80% chance that our statistical test will detect it.

The Figure illustrates the relationship between sample size and power for the test described. The orange line shows the power for the parameter settings above (baseline rate of 45% and

higher rates of smoking abstinence than varenicline alone.

The primary results were positive; that is, patients receiving

the combination therapy were more likely to achieve continuous

abstinence at 12 weeks than patients receiving varenicline alone.

The absolute difference in the abstinence rate was estimated to be

approximately 14%, which was statistically significant at level $\alpha = .05$.

These findings differed from the results reported in 2 previous

studies^{2,3} of the same question, which detected no difference in

treatments. What explains this difference? One explanation offered

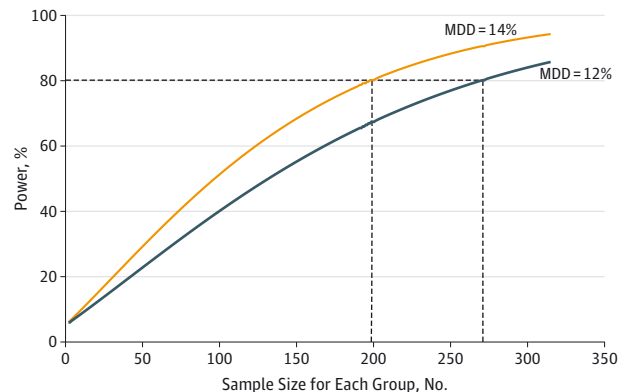
by the authors is that the previous studies "...may have been

inadequately powered," which means the sample size in those stud-

ies may have been too small to identify a difference between the

treatments tested.

Figure. Power for Detecting Difference and Sample Size



For a baseline rate of 45% and a minimum detectable difference (MDD) of 14%, the target sample size of 398 (199 in each group) will produce a power of 80% when α is set to .05. When the MDD is 12%, the resulting sample size is 542 (2×271) to achieve a power of 80%.

minimum detectable difference, or MDD, of 14%), when significance level α is set to .05. For this scenario, the authors' target sample size of 398 (199 in each group) will produce a power of 80%. All these values (45%, 14%, .05, 80%) must be selected at the planning stage of the study to carry out this calculation. The significance level and power are "rule-of-thumb" choices and are typically not based on the specifics of the study. If the researcher wants to reduce the probability of making a type I error ($\alpha = .05$) or to increase the probability of detecting the specified difference (power = 80%), then these values can be changed. Either change will require a larger sample size.

Selecting the baseline rate and MDD requires the expertise of the researcher. The baseline rate is typically available from the literature, because this rate is often based on a therapy that has been studied. The MDD choice is more subjective. It should be a clinically meaningful rate difference, or a scientifically important rate difference, or both, that is also feasible to detect. For example, if the combination therapy of varenicline and nicotine patch increased abstinence by 0.1%, this difference would not be of practical benefit, would require an extremely large sample size, and would thus be too small a setting for the MDD. If the MDD were specified as 50%, the new therapy would have to be 95% effective (45% + 50%) before there would be a high chance of detecting any difference, so would be too large for the MDD. The authors based their choice of MDD = 14% on a compromise between their judgment of a clinically important difference, 12%, and the scientifically meaningful value of 16%. The 16% rate was the observed difference in a study that compared varenicline alone and together with nicotine gum.⁴ Thus, the ability to con-

firm a difference that is slightly smaller for a related treatment was considered scientifically important.

What Are the Limitations of Power Analysis?

Calculation of sample size requires predictions of baseline rates and MDD, which may not be readily available, before the study begins. The sample size is especially sensitive to the MDD. This is illustrated by the blue line in the Figure, which shows the sample size needed in this study if the MDD were set to 12%. The resulting sample size is 542 (2×271) to achieve a power of 80%.

This method of conducting a power analysis might also produce the incorrect sample size if the data analysis conducted differs from that planned. For example, if abstinence were affected by other covariates, such as age, and the groups were unbalanced on this variable, other analyses might be used, such as logistic regression models accounting for covariate differences. The sample size that would be appropriate for one analysis may be too large or small to achieve the same power with another analytic procedure.

Why Did the Authors Use Power Analysis in This Particular Study?

The number of research participants available for any study is limited by resources. However, the authors were aware that previous studies comparing these treatments had found no significant difference in abstinence rates. This can occur even if a difference exists if the sample size is too small. The authors wanted to ensure that their sample size was adequate to detect even a small but clinically important difference, so they carefully evaluated sample size.

How Should This Method's Findings Be Interpreted in This Particular Study?

A power analysis can help with the interpretation of study findings when statistically significant effects are not found. However, because the findings in the study by Koegelenberg et al¹ were statistically significant, interpretation of a lack of significance was unnecessary. If no statistically significant difference in abstinence rates had been found, the authors could have noted that, "The study was sufficiently powered to have a high chance of detecting a difference of 14% in abstinence rates. Thus, any undetected difference is likely to be of little clinical benefit."

Caveats to Consider When Looking at Results Based on Power Analysis

Sample size calculation based on any power analysis requires input from the researcher prior to the study. Some of these are assumptions and predictions of fact (such as the baseline rate), which may be incorrect. Others reflect the clinical judgment of the researcher (eg, MDD), with which the reader may disagree. If a statistically significant effect is not found, the reader should assess whether either of these are concerns.

The reader should also not interpret a lack of significance for an outcome other than the one on which the power analysis was based as confirmation that no difference exists, because the analysis is specific to the parameter settings. For example, no significant difference was found in this study for most adverse events rates, although the power analysis does not apply to these rates. Thus, the sample size may not be adequate to interpret that finding to confirm that no meaningful difference in these outcomes exists.

ARTICLE INFORMATION

Author Affiliation: Department of Statistical Science, Southern Methodist University, Dallas, Texas.

Corresponding Author: Lynne Stokes, PhD, Department of Statistical Science, Southern Methodist University, PO Box 750100, Dallas, TX 75275 (slstokes@smu.edu).

Conflict of Interest Disclosures: The author has completed and submitted the ICMJE Form for

Disclosure of Potential Conflicts of Interest and none were reported.

REFERENCES

1. Koegelenberg CFN, Noor F, Bateman ED, et al. Efficacy of varenicline combined with nicotine replacement therapy vs varenicline alone for smoking cessation: a randomized clinical trial. *JAMA*. doi:10.1001/jama.2014.7195.
2. Hajek P, Smith KM, Dhanji AR, McRobbie H. Is a combination of varenicline and nicotine patch more effective in helping smokers quit than varenicline alone? a randomised controlled trial. *BMC Med*. 2013;11:140.
3. Ebbert JO, Burke MV, Hays JT, Hurt RD. Combination treatment with varenicline and nicotine replacement therapy. *Nicotine Tob Res*. 2009;11(5):572-576.
4. Besada NA, Guerrero AC, Fernandez MI, Ulibarri MM, Jiménez-Ruiz CA. Clinical experience from a smokers clinic combining varenicline and nicotine gum. *Eur Respir J*. 2010;36(suppl 54):462s.

JAMA Guide to Statistics and Methods

Interpretation of Clinical Trials That Stopped Early

Kert Viele, PhD; Anna McGlothlin, PhD; Kristine Broglio, MS

Clinical trials require significant resources to complete in terms of patients, investigators, and time and should be carefully designed and conducted so that they use the minimum amount of resources necessary to answer the motivating clinical question. The size of a clinical trial is typically based on the minimum number of patients required to have high probability of detecting the anticipated treatment effect. However, it is possible that strong evidence could emerge earlier in the trial either in favor of or against the benefit of the novel treatment. If early trial results are compelling, stopping the trial before the maximum planned sample size is reached presents ethical advantages for patients inside and outside the trial and can save resources that can be redirected to other clinical questions. This advantage must be balanced against the potential for overestimation of the treatment effect and other limitations of smaller trials (eg, limited safety data, less information about treatment effects in subgroups).

Many methods have been proposed to allow formal incorporation of early stopping into a clinical trial.^{1,2} All of these methods allow a trial to stop at a prespecified interim analysis while maintaining good statistical properties. Data monitoring committees or other similar governing bodies may also monitor the progress of a trial and recommend stopping the trial early in the absence of a prespecified formal rule. An overwhelmingly positive treatment effect might lead to a recommendation for unplanned early stopping but, more commonly, unplanned early stopping results from concerns for participant safety, lack of observed benefit, or concerns about the feasibility of continuing the trial due to slow patient accrual or new external information. Trials stopped for success in an ad hoc manner are challenging to interpret rigorously. In this article, we focus on early stopping for success or futility based on formal, prespecified stopping rules.

In the December 15, 2015, issue of *JAMA*, Stupp et al³ reported the results of a trial assessing electric tumor-treating fields plus temozolomide vs temozolomide alone in patients with glioblastoma. The trial design included a preplanned interim analysis defined according to an early stopping procedure. The trial was stopped for success at the interim analysis, reporting a hazard ratio of 0.62 for the primary end point of progression-free survival.

Use of the Method

Why Is Early Stopping Used?

When 2 treatments are compared in a randomized clinical trial, the treatment effects observed both during the trial and when the trial ends are subject to random highs and lows that depart from the true treatment effect. Sample sizes for trials are selected to reliably detect an anticipated treatment effect even if a modest, random low observed treatment effect occurs at the final analysis. If such a random low value does not occur or the true treatment effect is larger than anticipated, the extra study participants required to provide this protection against a false-negative result may not be neces-

sary. During the course of a trial, strong evidence may accumulate that the experimental treatment offers a benefit. This may be from a large observed treatment effect emerging early in a trial or from the anticipated treatment effect being observed as early as two-thirds of the way through a trial.

Conversely, evidence could accumulate early in a trial that the experimental treatment performs no better than the control. In a trial with no provision for early stopping, patients would continue to be exposed to the potential harms of the experimental therapy with no hope of benefit. Interim analyses to stop trials early for futility may avoid this risk. Trials may also stop early for futility if there is a limited likelihood of eventual success.⁴

What Are the Limitations of Early Stopping?

One key statistical issue with early stopping, particularly early stopping for success, is accounting for multiple "looks" at the data. Accumulating data, particularly early in the trial with a smaller number of observations, is likely to exhibit larger random highs and lows of values for the treatment effects. The more frequently the data are analyzed as they accumulate, the greater the chance of observing one of these fluctuations. Rules allowing early stopping therefore require a higher level of evidence, such as a lower *P* value, at each interim analysis than would be required at the end of a trial with no potential for early stopping. Taken together, the multiple looks at the data, each requiring a higher bar for success, lead to the same overall chance of falsely declaring success (type I error) as a trial with the usual criterion for success (eg, a $P < .05$) and no potential for early stopping.

Early stopping for futility requires no such adjustment. There are no added opportunities to declare a success; thus, no statistical adjustment to the success threshold is required. However, futility stopping may reduce the power of the trial by stopping trials based on a random low value for the treatment effect that could have gone on to be successful. This reduction in power is usually quite small.

Success thresholds are typically chosen to be more conservative for interim analyses than for the final analysis should the trial continue to completion. The O'Brien-Fleming method, for example, requires very small *P* values to declare success early in the trial and then maintains a final *P* value very close to the traditional .05 level at the final analysis.¹ Using this method, very few trials could be successful at the interim analyses that would not have been successful at the final analysis. Thus, there is a minimal "penalty" for the interim analyses. The more conservative the early stopping criteria, the more assurance there is that an early stop for success is not a false-positive result.

While methods such as O'Brien-Fleming protect against falsely declaring an ineffective drug successful, the accuracy of estimates of the treatment effect in trials that have stopped early for success remains a concern.⁵ When considering the true effect of a treatment, bias is introduced when considering only trials that have observed a large enough treatment effect to meet the critical value for

success. By definition, successful trials have larger treatment effects than unsuccessful trials; thus, successful trials include more random highs than random lows. As such, small trials that end in success, either at the end or early, are prone to overestimating the treatment effect. The larger the observed treatment effect, the more likely it is an extreme random high, and the greater the chance for overestimation. If the trial were continued, with the enrollment of additional patients, it is likely that there would be a reduction of the observed treatment effect. In other words, trials with very impressive early results are likely to become less impressive after observing more data, and this should be taken into account when monitoring and interpreting such trials. Extreme attenuation, such as a complete disappearance of the observed treatment benefit, however, is less likely.

Why Did the Authors Use Early Stopping in This Study?

Glioblastoma is an aggressive cancer with few treatment options. In the report by Stupp et al,³ enrollment was largely complete at the time of the interim analysis. However, the interim analysis allowed the possibility that a beneficial result could be disseminated many months (potentially years) earlier in advance of the fully mature data.

How Should Early Stopping Be Interpreted in This Particular Study?

The primary analysis in this study found a hazard ratio of 0.62 ($P = .001$) based on 18 months of follow-up from the first 315 patients enrolled. This is strong evidence of a treatment benefit for tumor-treating fields plus temozolomide in this population. However, care should be taken when interpreting the estimated benefit

corresponding to a hazard ratio of 0.62. Given the potential for an overestimated treatment effect, combined with the general intractability of treating glioblastoma, there is good reason to suspect that the actual benefit of tumor-treating fields, while present, might be smaller than that observed in the study. A robustness analysis (ie, a supplementary or supporting analysis conducted to see how consistent the results are if different approaches were taken in conducting the analysis), based on the then-available data from all participants, illustrates this pattern. That analysis resulted in a hazard ratio of 0.69 (95% CI, 0.55-0.86), also with a $P < .001$. The result remained statistically significant, but the magnitude of the treatment effect was smaller.

Caveats to Consider When Looking at a Trial That Stopped Early

It is important to consider trial design, quality of trial conduct, safety and secondary end points, and other supplementary data when interpreting the results of any clinical trial. For trials that stop early for success, the statistical superiority of an experimental treatment is straightforward when the early stopping was preplanned and it is reasonable to preserve patient resources and time once the primary objective of a trial has been addressed. Early stopping procedures protect against a false conclusion of superiority. However, if the result seems implausibly good, there is a high likelihood that the true effect is smaller than the observed effect. In that light, the benefits of early stopping, to patients both in and out of the trial, must be weighed against how much potential additional knowledge would be gained if the trial were continued.

ARTICLE INFORMATION

Author Affiliations: Berry Consultants LLC, Austin, Texas.

Corresponding Author: Kert Viele, PhD, Berry Consultants LLC, 4301 Westbank Dr, Bldg B, Ste 140, Austin, TX 78746 (kert@berryconsultants.com).

Section Editors: Roger J. Lewis, MD, PhD, Department of Emergency Medicine, Harbor-UCLA Medical Center and David Geffen School of Medicine at UCLA; and Edward H. Livingston, MD, Deputy Editor, *JAMA*.

Conflict of Interest Disclosures: All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

REFERENCES

1. Jennison C, Turnbull BW. *Group Sequential Methods With Applications to Clinical Trials*. Boca Raton, FL: Chapman & Hall; 2000.
2. Broglio KR, Connor JT, Berry SM. Not too big, not too small: a Goldilocks approach to sample size selection. *J Biopharm Stat*. 2014;24(3):685-705.
3. Stupp R, Taillibert S, Kanner AA, et al. Maintenance therapy with tumor-treating fields plus temozolomide vs temozolomide alone for glioblastoma: a randomized clinical trial. *JAMA*. 2015;314(23):2535-2543.
4. Saville BR, Connor JT, Ayers GD, Alvarez J. The utility of Bayesian predictive probabilities for interim monitoring of clinical trials. *Clin Trials*. 2014;11(4):485-493.
5. Zhang JJ, Blumenthal GM, He K, Tang S, Cortazar P, Sridhara R. Overestimation of the effect size in group sequential trials. *Clin Cancer Res*. 2012;18(18):4872-4876.

JAMA Guide to Statistics and Methods

Cluster Randomized Trials

Evaluating Treatments Applied to Groups

William J. Meurer, MD, MS; Roger J. Lewis, MD, PhD

Sometimes a new treatment is best introduced to an entire group of patients rather than to individual patients. Examples include when the new approach requires procedures be followed by multiple members of a health care team or when the new technique is applied to the environment of care (eg, a method for cleaning a hospital room before it is known which patient will be assigned the room). This avoids confusion that could occur if all caregivers had to keep track of which patients were being treated the old way and which were being treated the new way.

One approach to evaluate the efficacy of such treatments—treatments for which the application typically involves changes at the level of the health care practice, hospital unit, or even health care system—is to conduct a cluster randomized trial. In a cluster randomized trial, study participants are randomized in groups or clusters so that all members within a single group are assigned to either the experimental intervention or the control.^{1,2} This contrasts with the more familiar randomized clinical trial (RCT) in which randomization occurs at the level of the individual participant, and the treatment assigned to one study participant is essentially independent of the treatment assigned to any other. In a cluster randomized trial, the cluster is the unit randomized, whereas in a traditional RCT, the individual study participant is randomized. In both types of trials, however, the outcomes of interest are recorded for each participant individually.

Although there are both theoretical and pragmatic reasons for using cluster randomization in a clinical trial, doing so introduces a fundamental challenge to those analyzing and interpreting the results of the trial: study participants from the same cluster (eg, patients treated within the same medical practice or hospital unit) tend to be more similar to each other than participants from different clusters.² This nearly universal fact violates a common assumption of most statistical tests, namely, that individual observations are independent of each other. To obtain valid results, a cluster randomized trial must be analyzed using statistical methods that account for the greater similarity between individual participants from the same cluster compared with those from different clusters.²⁻⁴

In a recent *JAMA* article, Curley et al⁵ reported the results of RESTORE, a cluster randomized clinical trial evaluating a nurse-implemented, goal-directed sedation protocol for children with acute respiratory failure receiving mechanical ventilation in the intensive care setting, comparing this approach with usual care. The trial evaluated the primary hypothesis that the intervention group—patients treated in intensive care units (ICUs) using the goal-directed sedation protocol—would have a shorter duration of mechanical ventilation. Thirty-one pediatric ICUs, the “clusters,” were randomized to either implement the goal-directed sedation protocol or continue their usual care practices.

Use of the Method

Why Is Cluster Randomization Used?

Cluster randomization should be used when it would be impractical or impossible to assign and correctly deliver the experimental and control treatments to individual study participants.^{1,2} Typical situations include the study of interventions that must be implemented by multiple team members, that affect workflow, or that alter the structure of care delivery. As in the RESTORE trial, interventions that involve training multidisciplinary health care teams are practically difficult to conduct using individual-level randomization, as health care practitioners cannot easily unlearn a new way of taking care of patients.

Cluster randomization is often used to reduce the mixing or contamination of treatments in the 2 groups of the trial, as might occur if patients in the control group start to be treated using some of the approaches included in the experimental treatment group, perhaps because the practitioners become habituated to the experimental approach or perceive it to be superior.^{1,2} For example, consider an injury prevention trial testing the effect of offering bicycle helmets to students in a classroom on the incidence of subsequent head injury. If a conventional RCT were conducted and half of the students in each classroom received helmets, it is likely that some of the other half of students would inform their parents about the ongoing intervention and many of these children might also begin to use bicycle helmets. Contamination is a form of crossover between treatment groups and will generally reduce the observed treatment effect using the usual intent-to-treat analysis.⁶ Cluster randomization may also be used to reduce potential selection bias. Physicians choosing individual patients from their practice to consent for randomization may tend to enroll patients with specific characteristics (eg, lesser or greater illness severity), reducing the external validity of the trial. Assignment of the treatment group at the practice level, with the application of the assigned treatment to all patients treated within the practice, may minimize this problem.

Using a cluster randomized design also can offer practical advantages. For example, if 2 or more treatments are considered to be within the standard of care, and depending on the risks associated with treatment, streamlined consent procedures or even integration of general and research consents may be used to reduce barriers to participation and ensure a truly representative patient population is enrolled in the trial.^{1,7}

What Are Limitations of Cluster Randomization?

Any time data are clustered, the statistical analysis must use techniques that account for the likeness of cluster members.^{2,3} Extensions of the more-familiar regression models that are appropriate for the analysis of clustered data include generalized estimating equations (as used in RESTORE), mixed linear models, and hierarchical

models. While the proper use of these approaches is complex, the informed reader should be alert to statements that the analysis method was selected to account for the similarity or correlations of data within each cluster. The intraclass correlation coefficient (ICC) quantifies the likeness within clusters and ranges from 0 to 1, although it is frequently in the 0.02 to 0.1 range.⁴ A value of 0 means each member of the cluster is not more like the other members, with respect to the measured characteristic, than they are to the population at large, so each additional individual contributes the same amount of new information. In contrast, a value of 1 means that each member of the cluster is exactly the same as the others in the cluster, so any participants beyond the first contribute no additional information at all. A larger ICC, representing greater similarity of results within clusters, will decrease the effective sample size of the trial, reducing the precision of estimates of treatment effects and the power of the trial.² If the ICC is high, the effective sample size will be closer to the number of groups, and if the ICC is low, the effective sample size will be closer to the total number of individuals in the trial.

It is often impossible to maintain blinding of treatment assignment in a cluster randomized trial, both because of the nature of treatments and because of the number of patients in a given location all receiving the same treatment. It is well known that trials evaluating nonblinded interventions have a greater risk of bias.

Why Did the Authors Use Cluster Randomization in This Particular Study?

The RESTORE trial investigators used cluster randomization because they were introducing a nurse-implemented, goal-directed sedation protocol that required a change in behavior among multiple caregivers within each ICU. A major component of the experimental intervention was educating the critical care personnel regarding the perceived benefits and risks of sedation agents and use patterns relative to others. Had individual-level randomization been used to allocate patients, it is highly likely that the patients randomized to standard care would have received care that was somewhere between the prior standard and the new protocol, because all ICU caregivers would have been informed about the scientific and pharmacological basis for the goal-directed sedation protocol.

How Should Cluster Randomization Findings Be Interpreted in This Particular Study?

As in any clinical trial, randomization may or may not work effectively to create similar groups of patients. In RESTORE, some differences between the intervention groups were observed that might partially explain the negative primary outcome. Specifically, the intervention group had a greater proportion of younger children—a group that is more difficult to sedate.⁸ The RESTORE investigators used randomization in blocks to ensure balance of pediatric ICU sizes between groups; methods exist to balance groups in cluster trials on multiple factors simultaneously.⁹ Although the RESTORE trial yielded a negative primary outcome, the authors noted some promising secondary outcomes related to clinicians' perception of patient comfort. However, these assessments were unblinded and thus may be subject to bias.

Caveats to Consider When Looking at a Cluster Randomized Trial

When evaluating a cluster randomized trial, the first consideration is whether the use of clustering was well justified. Would it have been possible to use individual-level randomization and maintain fidelity in treatment allocation and administration? What would be the likelihood of contamination? Cluster randomization cannot minimize baseline differences between 2 treatment groups as efficiently as individual-level randomization. The design must be justified for scientific or logistical reasons to accept this trade-off.¹⁰

Second, the usual sources of bias should be considered, such as patient knowledge of treatment assignment and unblinded assessments of outcome. Although not specific to cluster randomized trials, these sources of bias tend to be more problematic.

Third, it is important to consider whether the intraclass correlation was appropriately accounted for in the design, analysis, and interpretation of the trial.^{11,10} During the design, the likely ICC should be considered to ensure the planned sample size is adequate. The analysis should be based on statistical methods that account for clustering, such as generalized estimating equations.

Finally, the interpretation should consider the extent with which the 2 treatment groups contained an adequate number, size, and similarity of clusters and whether any clusters were lost to follow-up.

ARTICLE INFORMATION

Author Affiliations: Department of Emergency Medicine, University of Michigan, Ann Arbor (Meurer); Department of Emergency Medicine, Harbor-UCLA Medical Center, Torrance, California (Lewis); Los Angeles Biomedical Research Institute, Torrance, California (Lewis); David Geffen School of Medicine at University of California, Los Angeles (Lewis); Berry Consultants, Austin, Texas (Lewis).

Corresponding Author: Roger J. Lewis, MD, PhD, Department of Emergency Medicine, Harbor-UCLA Medical Center, 1000 W Carson St, Bldg D9, Torrance, CA 90509 (roger@emedharbor.edu).

Section Editors: Roger J. Lewis, MD, PhD, Department of Emergency Medicine, Harbor-UCLA Medical Center and David Geffen School of Medicine at UCLA; and Edward H. Livingston, MD, Deputy Editor, *JAMA*.

Conflict of Interest Disclosures: Both authors have completed and submitted the ICMJE Form for

Disclosure of Potential Conflicts of Interest and none were reported.

REFERENCES

- Campbell MK, Elbourne DR, Altman DG; CONSORT group. CONSORT statement: extension to cluster randomised trials. *BMJ*. 2004;328(7441):702-708.
- Wears RL. Advanced statistics: statistical methods for analyzing cluster and cluster-randomized data. *Acad Emerg Med*. 2002;9(4):330-341.
- Dawid AP. Conditional independence in statistical theory. *J R Stat Soc Series B*. 1979;41:1-31.
- Killip S, Mahfoud Z, Pearce K. What is an intraclass correlation coefficient? *Ann Fam Med*. 2004;2(3):204-208.
- Curley MA, Wypij D, Watson RS, et al. Protocolized sedation vs usual care in pediatric patients mechanically ventilated for acute respiratory failure. *JAMA*. 2015;313(4):379-389.
- Detry MA, Lewis RJ. The intention-to-treat principle. *JAMA*. 2014;312(1):85-86.
- Huang SS, Septimus E, Kleinman K, et al. Targeted versus universal decolonization to prevent ICU infection. *N Engl J Med*. 2013;368(24):2255-2265.
- Anand KJ, Willson DF, Berger J, et al. Tolerance and withdrawal from prolonged opioid use in critically ill children. *Pediatrics*. 2010;125(5):e1208-e1225.
- Scott PA, Meurer WJ, Frederiksen SM, et al. A multilevel intervention to increase community hospital use of alteplase for acute stroke (INSTINCT). *Lancet Neurol*. 2013;12(2):139-148.
- Ivers NM, Taljaard M, Dixon S, et al. Impact of CONSORT extension for cluster randomised trials on quality of reporting and study methodology. *BMJ*. 2011;343:d5886.

Case-Control Studies

Using “Real-world” Evidence to Assess Association

Telba Z. Irony, PhD

Associations between patient characteristics or treatments received and clinical outcomes are often first described using observational data, such as data arising through usual clinical care without the experimental assignment of treatments that occurs in a randomized clinical trial (RCT). These data based on usual clinical care are referred to by some as “real-world” data. A key strategy for efficiently finding such associations is to use a case-control study.¹ In a recent issue of *JAMA Internal Medicine*, Wang et al² assessed the association between cardiovascular disease (CVD) and use of inhaled long-acting β_2 -agonists (LABAs) or long-acting antimuscarinic antagonists (LAMAs) in chronic obstructive pulmonary disease (COPD), utilizing a nested case-control study.

Explanation of the Method

What Are Case-Control and Nested Case-Control Studies?

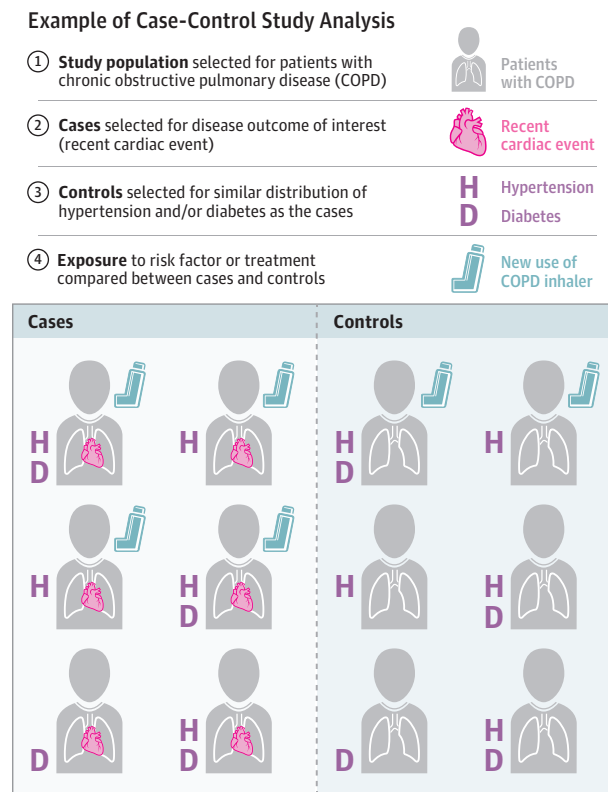
A case-control study compares individuals who had the outcome of interest (cases) vs individuals who did not have that outcome (controls) with respect to exposure to a potential “risk factor.” The goal is to determine if there is an association between the risk factor and the outcome. The risk factor may be a behavior such as tobacco use, a patient characteristic, or a treatment. The idea is to define a population or cohort, identify the cases and controls in the population, and retrospectively determine which patients in each group were exposed to the risk factor; the case-control study works backward from outcome to exposure (Figure). A higher proportion of individuals with exposure to the risk factor among cases than among controls suggests that the risk factor is associated with the outcome. The term *control* refers to an individual who did not have the outcome; in contrast, the same term in a clinical trial refers to a study participant who receives the standard (or placebo) treatment.

In a nested case-control study, the cases are identified in a large cohort and, for each case, a specified number of controls matching the case are selected from the cohort. The selected controls should match the cases with respect to characteristics, other than the risk factor, that are likely related to the outcome of interest. Because it is easier to find controls than cases when the outcome is rare, increasing the number of controls beyond the number of cases (eg, 2:1 or 3:1 matching) may be used to improve study precision.

The nested case-control study by Wang et al² used data from 284 220 LABA-LAMA-naïve patients with COPD retrieved from the Taiwan National Health Insurance Research Database with health care claims from 2007 to 2011. Cases (n = 37 719) were patients who had inpatient or emergency care visits for coronary artery disease, heart failure, ischemic stroke, or arrhythmia (CVD events). Each patient was matched to 4 controls (n = 146 139) without visits for these disorders.

In a case-control study, the most common measure of association between exposure and outcome is the odds ratio (OR), which aims to compare the occurrence of the outcome in the presence of the expo-

Figure. Hypothetical Example of a Case-Control Study



Exposure to a risk factor (in this case, new COPD inhaler use) changes the chance of subsequently developing the outcome of interest. However, in conducting a case-control study, the outcome (in this case, a cardiovascular event) is used initially to define cases and controls, and then the distribution of the exposure is assessed.

sure vs in the absence of the exposure.³ In practice, the OR in a case-control study is the ratio of the odds of exposure among the cases to the odds of exposure among the controls, where the odds of exposure is the probability of exposure divided by the probability of no exposure. The prevalence of the exposure is compared between cases and controls and not the other way around. However, because the OR treats outcome and exposure symmetrically, it provides the desired measure of association. If the OR is greater than 1, the exposure is associated with the outcome, ie, having the exposure increases the odds of having an outcome (and vice versa). The OR is a measure of effect size; the larger the OR, the stronger the association.

In the study by Wang et al,² new use of LABA occurred in 520 cases (1.4%) and 1186 controls (0.8%), resulting in an adjusted odds ratio of 1.50 (95% CI, 1.35-1.67). New use of LAMA occurred in 190 cases (0.5%) and 463 controls (0.3%), resulting in an adjusted odds ratio of 1.52 (95% CI, 1.28-1.80). An OR of 1.5 represents a modest association⁴ between outcome (CVD) and exposure (LABA and

LAMA). Thus, the authors found that new use of LABAs or LAMAs was associated with a modest increase in cardiovascular risk in patients with COPD, within 30 days of therapy initiation.

Why Are Case-Control Studies Used?

Case-control studies are time-efficient and less costly than RCTs, particularly when the outcome of interest is rare or takes a long time to occur, because the cases are identified at study onset and the outcomes have already occurred with no need for a long-term follow-up. The case-control design is useful in exploratory studies to assess a possible association between an exposure and outcome. Nested case-control studies are less expensive than full cohort studies because the exposure is only assessed for the cases and for the selected controls, not for the full cohort.

Limitations of Case-Control Studies

Case-control studies are retrospective and data quality must be carefully evaluated to avoid bias. For instance, because individuals included in the study and evaluators need to consider exposures and outcomes that happened in the past, these studies may be subject to recall bias and observer bias. Because the controls are selected retrospectively, such studies are also subject to selection bias, which may make the case and control groups not comparable. For a valid comparison, appropriate controls must be used, ie, selected controls must be representative of the population that produced the cases. The ideal control group would be generated by a random sample from the general population that generated the cases. If controls are not representative of the population, selection bias may occur.

Case-control studies provide less compelling evidence than RCTs. Due to randomization, treatment and control groups in RCTs tend to be similar with respect to baseline variables, including unmeasured ones.⁵ Because the only difference between treatment and control groups is the treatment, RCTs can demonstrate causation between treatment and outcome. In case-control studies, case and control groups are similar with respect to the matching variables, but are not necessarily similar with respect to unmeasured variables. Such studies are susceptible to confounding, which occurs when the exposure and the outcome are both associated with a third unmeasured variable.⁶ Unlike RCTs, case-control studies demonstrate association between exposure and outcome but do not demonstrate causation.

The objective of case-control studies is to compare the occurrence of an outcome with and without an exposure. The relative risk

(RR), which is the ratio between the probability of the outcome when exposed and the probability of the outcome when not exposed, provides a straightforward comparison measure but, because the case-control study design does not allow for the estimation of the occurrence of the outcome in the population (ie, incidence or prevalence), the RR cannot be determined from a case-control study. A case-control study can only estimate the OR, which is the ratio of odds and not the ratio of probabilities. The OR approximates the RR for rare outcomes, but differs substantially when the outcome of interest is common. In addition, case-control studies are limited to the examination of one outcome, and it is difficult to examine the temporal sequence between exposure and outcome.

Despite these limitations, case-control studies and other "real-world" evidence can provide valuable empirical evidence to complement RCTs. Additionally, case-control studies may be able to address questions for which an RCT is either not feasible or not ethical.⁷

How Was the Method Applied in This Case?

In the case-control study by Wang et al,² the exposure to LABA and LAMA use for both cases and controls in the year preceding the occurrence of the CVD event was measured and stratified by duration since initiation of LABA or LAMA into 4 groups: current (≤ 30 days), recent (31-90 days), past (91-180 days), and remote (>180 days). Additional stratification on concomitant COPD medications and other factors was also conducted. The data source used in the study (Taiwan National Health Insurance Research Database) mitigates data quality concerns because it is national, universal, compulsory, and subject to periodic audits. Overall, the authors found that new use of LABAs or LAMAs was associated with a modest increase in cardiovascular risk in patients with COPD, within 30 days of therapy initiation, and this finding was strengthened by the steps taken to ensure data quality and comparability of cases and controls.

How Does the Case-Control Design Affect the Interpretation of the Study?

Causality cannot be established in a case-control study because there is no way to control for unmeasured confounders. In the study by Wang et al,² the use of the disease risk score for predicting CVD events was helpful to control for measured confounders but could not adjust for unmeasured confounders. The authors mitigated further possible confounding effects by conducting extensive sensitivity analyses.

ARTICLE INFORMATION

Author Affiliation: Office of Biostatistics and Epidemiology, Center for Biologics Evaluation and Research (CBER), US Food and Drug Administration, Silver Spring, Maryland.

Corresponding Author: Telba Z. Irony, PhD, Office of Biostatistics and Epidemiology, CBER/FDA, 10903 New Hampshire Ave, Bldg 71, 1216, Silver Spring, MD 20953 (telba.irony@fda.hhs.gov).

Published Online: August 23, 2018.
doi:10.1001/jama.2018.12115

Conflict of Interest Disclosures: The author has completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

Disclaimer: This article reflects the views of the author and should not be construed to represent FDA's views or policies.

REFERENCES

- Breslow NE. Statistics in epidemiology: the case-control study. *J Am Stat Assoc*. 1996;91(433):14-28. doi:10.1080/01621459.1996.10476660
- Wang MT, Liou JT, Lin CW, et al. Association of cardiovascular risk with inhaled long-acting bronchodilators in patients with chronic obstructive pulmonary disease: a nested case-control study. *JAMA Intern Med*. 2018;178(2):229-238. doi:10.1001/jamainternmed.2017.7720
- Norton EC, Dowd BE, Maciejewski ML. Odds ratios: current best practice and use. *JAMA*. 2018;320(1):84-85. doi:10.1001/jama.2018.6971
- Chen H, Cohen P, Chen S. How big is a big odds ratio? interpreting the magnitudes of odds ratios in epidemiological studies. *Commun Stat Simul Comput*. 2010;39(4):860-864.
- Broglio K. Randomization in clinical trials: permuted blocks and stratification. *JAMA*. 2018;319(21):2223-2224. doi:10.1001/jama.2018.6360
- Kyriacou DN, Lewis RJ. Confounding by indication in clinical research. *JAMA*. 2016;316(17):1818-1819. doi:10.1001/jama.2016.16435
- Corrigan-Curay J, Sacks L, Woodcock J. Real-world evidence and real-world data for evaluating drug safety and effectiveness [published online August 13, 2018]. *JAMA*. 2018. doi:10.1001/jama.2018.10136

JAMA Guide to Statistics and Methods

Decision Curve Analysis

Mark Fitzgerald, PhD; Benjamin R. Saville, PhD; Roger J. Lewis, MD, PhD

Decision curve analysis (DCA) is a method for evaluating the benefits of a diagnostic test across a range of patient preferences for accepting risk of undertreatment and overtreatment to facilitate



Related article page 390

decisions about test selection and use.¹ In this issue of *JAMA*, Siddiqui and colleagues² used DCA to evaluate 3 prostate

biopsy strategies: targeted magnetic resonance/ultrasound fusion biopsy, standard extended-sextant biopsy, or a combination, for establishing the diagnosis of intermediate- to high-risk prostate cancer. Their goal was to identify the best biopsy strategy to ensure prostatectomy is offered to patients with intermediate- and high-risk tumors and avoided for patients with low-risk tumors.

Use of the Method

Why Is DCA Used?

When patients have signs or symptoms suggestive of but not diagnostic of a disease, they and their physician must decide whether to (1) treat empirically, (2) not treat, or (3) perform further diagnostic testing before deciding between options 1 and 2. The decision to treat depends on how confident the clinician is that the disease is present, the effectiveness and complications of treatment if the disease is present, and the patient's willingness to accept the risks and burden of a treatment that might not be necessary. A diagnostic test may provide additional information on whether the disease is present.³ Decision curve analysis is a method to assess the value of information provided by a diagnostic test by considering the likely range of a patient's risk and benefit preferences, without the need for actually measuring these preferences for a particular patient.¹

A key concept in DCA is that of a "probability threshold," namely, a level of diagnostic certainty above which the patient would choose to be treated. The probability threshold used in DCA captures the relative value the patient places on receiving treatment for the disease, if present, to the value of avoiding treatment if the disease is not present. If the treatment has high efficacy and minimal cost, inconvenience, and adverse effects (eg, oral antibiotics for community-acquired pneumonia), then the probability threshold will be low; conversely, if the treatment is minimally effective or associated with substantial morbidity (eg, radiation for a malignant brain tumor), then the probability threshold will be high.

The net benefit, or "benefit score," is determined by calculating the difference between the expected benefit and the expected harm associated with each proposed testing and treatment strategy. The expected benefit is represented by the number of patients who have the disease and who will receive treatment (true positives) using the proposed strategy.

The expected harm is represented by number of patients without the disease who would be treated in error (false positives) multiplied by a weighting factor based on the patient's threshold probability. The weighting factor captures the patient's values regarding

the risks of undertreatment and overtreatment. Specifically, the false-positive rate is multiplied by the ratio of the threshold probability divided by $1 -$ the threshold probability. For example, if the treatment threshold is 10% (0.1) for a patient with possible pneumonia, then the weighting factor applied to the number of patients without pneumonia treated in error would be $0.1/0.9$, or one-ninth, minimizing the effect of false-positive results because the burden of unnecessary treatment is low. Conversely, for a patient with a brain mass that is possibly malignant, the probability threshold might be 90% (0.9), leading to a weighting factor of $0.9/0.1$, or 9, and greatly increasing the effect of the risk of false-positive results with any proposed testing and treatment strategy.

Graphically, the DCA is expressed as a curve, with benefit score on the vertical axis and probability thresholds on the horizontal axis. A curve is drawn for each approach that might be taken to establish a diagnosis. Another line is drawn to show what happens when no treatment is ever given (ie, no net benefit), and another curve is drawn as if all patients receive treatment irrespective of test results. For any given patient's probability threshold, the curve with the highest benefit score at that threshold is the best choice.¹

If one curve is highest over the full range of probability thresholds, then the associated diagnostic approach would be the best decision for all patients, regardless of individual values, and a clinician can use this approach uniformly. If the curves cross, then the optimal approach will depend on the patient's risk tolerance, expressed through their probability threshold.

What Are the Limitations of the DCA Method?

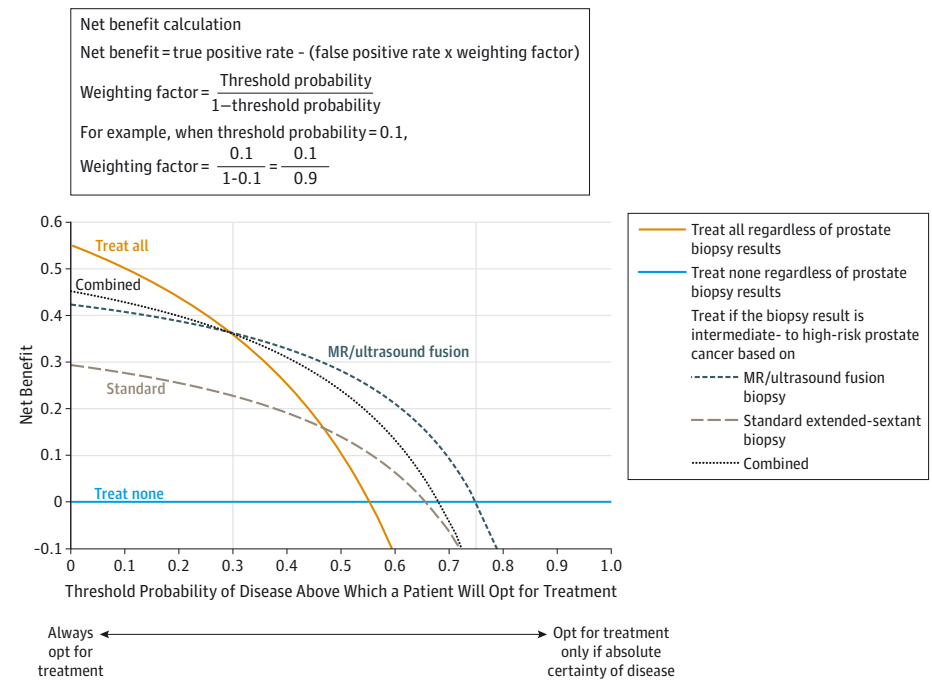
For diseases that are not well studied, there may be insufficient knowledge regarding patient preferences to determine the relevant range of threshold probabilities. Even when the likely range of probability thresholds is known, if the decision curves cross within that range, then the clinician must delve deeper into individual patient preferences to choose a testing and treatment strategy.³

Caution should be used in interpreting DCAs based on published ranges of threshold probabilities, particularly when there are many treatment options available to a patient. A patient is likely to have a different threshold probability if treatment is, for example, radiation rather than prostatectomy. The threshold probability needs to apply to a well-defined path of treatment.

Decision curve analysis does not explicitly account for the costs (monetary costs, time lost, physical or psychological discomfort, etc) associated with the diagnostic test. Further, if the diagnostic test provides information about how to treat as well as whether to treat (eg, a biopsy that yields both a cancer diagnosis and tumor type, allowing the selection of a specific therapy), the decision curve does not incorporate the value of this additional information.

Another challenge in correct implementation of DCA is that the data required for establishing the curve are often difficult to obtain. There must be sufficient study data for the population of in-

Figure. Net Benefit as a Function of a Threshold Probability of Intermediate- to High-Risk Prostate Cancer



Threshold probability refers to the point at which a patient considers the benefit of treatment for intermediate- to high-risk prostate cancer equivalent to the harm of overtreatment for low-risk disease and thus reflects how the patient weights the benefits and harms associated with this decision. The highest curve at any given threshold probability is the optimal decision-making strategy to maximize net benefit. Net benefit was maximized with threshold probabilities of 0%-30% by the "treat all" approach; with threshold probabilities of 30%-75%, net benefit was maximized by the targeted magnetic resonance (MR)/ultrasound fusion approach; and with 75%-100%, net benefit was maximized by the "treat none" approach. (Adapted from Supplement, Siddiqui et al.²)

terest to whom the diagnostic test has been applied and the true state of the disease known for each patient at the time of the test. A fairly large patient study may be needed to establish estimates of traditional measures of accuracy (sensitivity, specificity).

Why Did the Authors Use DCA in This Particular Study?

There is controversy surrounding the benefits of screening and intervention relative to the costs of unnecessarily treating low-risk prostate cancers.^{4,5} Justification for use of MR/ultrasound fusion-guided biopsy with ultrasound-guided biopsy to diagnose prostate cancer must be shown to benefit a broad range of patients.

How Should DCA Findings Be Interpreted in This Particular Study?

The DCA reported by Siddiqui et al² showed that for patients with threshold probabilities of 0% to 30%, representing a relative preference for empirical treatment, the net benefit is greatest if all patients are treated and that the diagnostic tests do not add sufficient information to improve care (Figure). In this range of threshold probabilities, patients appear to be more concerned about missing

a diagnosis of cancer than about receiving unnecessary treatment. For midrange threshold probabilities of 30% to 75%, the targeted biopsy approach is superior to other strategies, including the 2 other diagnostic approaches evaluated. For higher thresholds (>75%) at which patients may be more concerned about unnecessary treatment than missed cancer, the option to not treat is preferred and neither diagnostic test has value.

Caveats to Consider When Looking at Results Based on DCA

One shortcoming of this study was the use of a subset of 170 patients who underwent prostatectomy in constructing the DCA. These patients self-selected for prostatectomy after learning the results of their targeted and standard biopsies. This group primarily comprised men who had higher cancer risk, resulting in potential bias when estimating false positives, false negatives, and other diagnostic measures. The patients classified as low risk who still opted for prostatectomy are patients with low probability thresholds, who might also be different from the broader population of men with symptoms or findings suggesting prostate cancer.

ARTICLE INFORMATION

Author Affiliations: Berry Consultants, Austin, Texas (Fitzgerald, Saville, Lewis); Department of Biostatistics, Vanderbilt University School of Medicine, Nashville, Tennessee (Saville); Department of Emergency Medicine, Harbor-UCLA Medical Center, Torrance, California (Lewis); Los Angeles Biomedical Research Institute, Torrance, California (Lewis); David Geffen School of Medicine, University of California, Los Angeles (Lewis).

Corresponding Author: Roger J. Lewis, MD, PhD, Department of Emergency Medicine, Harbor-UCLA Medical Center, 1000 W Carson St, Bldg D9, Torrance, CA 90509 (roger@emedharbor.edu).

Conflict of Interest Disclosures: All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

REFERENCES

- Vickers AJ, Elkin EB. Decision curve analysis. *Med Decis Making*. 2006;26(6):565-574.
- Siddiqui MM, Rais-Bahrami S, Turkbey B, et al. Comparison of MR/ultrasound fusion-guided biopsy with ultrasound-guided biopsy for the diagnosis of prostate cancer. *JAMA*. doi:10.1001/jama.2014.17942.

- Sox HC, Higgins MC, Owens DK. *Medical Decision Making*. 2nd ed. West Sussex, UK: John Wiley & Sons; 2013.

- Froberg DG, Kane RL. Methodology for measuring health-state preferences. *J Clin Epidemiol*. 1989;42(4-7):345-354.

- Hoffman RM. Clinical practice: screening for prostate cancer. *N Engl J Med*. 2011;365(21):2013-2019.

JAMA Guide to Statistics and Methods

Gatekeeping Strategies for Avoiding False-Positive Results in Clinical Trials With Many Comparisons

Kabir Yadav, MDCM, MS, MSHS; Roger J. Lewis, MD, PhD

Clinical trials characterizing the effects of an experimental therapy rarely have only a single outcome of interest. In a previous report in *JAMA*,¹ the CLEAN-TAVI investigators evaluated the benefits of a cerebral embolic protection device for stroke prevention during transcatheter aortic valve implantation. The primary end point was the reduction in the number of ischemic lesions observed 2 days after the procedure. The investigators were also interested in 16 secondary end points involving measurement of the number, volume, and timing of cerebral lesions in various brain regions. Statistically comparing a large number of outcomes using the usual significance threshold of .05 is likely to be misleading because there is a high risk of falsely concluding that a significant effect is present when none exists.² If 17 comparisons are made when there is no true treatment effect, each comparison has a 5% chance of falsely concluding that an observed difference exists, leading to a 58% chance of falsely concluding at least 1 difference exists. The formula $1 - [1 - \alpha]^N$ can be used to calculate the chance of obtaining at least 1 falsely significant result, when there is no true underlying difference between the groups (in this case α is .05 and N is 17 for the number of tests).

To avoid a false-positive result, while still comparing the multiple clinically relevant end points used in the CLEAN-TAVI study, the investigators used a serial gatekeeping approach for statistical testing. This method tests an outcome, and if that outcome is statistically significant, then the next outcome is tested. This minimizes the chance of falsely concluding a difference exists when it does not.

Use of the Method

Why Is Serial Gatekeeping Used?

Many methods exist for conducting multiple comparisons while keeping the overall trial-level risk of a false-positive error at an acceptable level. The Bonferroni approach³ requires a more stringent criterion for statistical significance (a smaller P value) for each statistical test, but each is interpreted independently of the other comparisons. This approach is often considered to be too conservative, reducing the ability of the trial to detect true benefits when they exist.⁴ Other methods leverage additional knowledge about the trial design to allow only the comparisons of interest. In the Dunnett method for comparing multiple experimental drug doses against a single control, the number of comparisons is reduced by never comparing experimental drug doses against each other.⁵ Multiple comparison procedures, including the Hochberg procedure, have been discussed in a prior *JAMA* Guide to Statistics and Methods.²

Description of the Method

A serial gatekeeping procedure controls the false-positive risk by requiring the multiple end points to be compared in a predefined sequence and stopping all further testing once a nonsignificant result is obtained. A given comparison might be considered positive if it were placed early in the sequence, but the same analysis would be considered negative

if it were positioned in the sequence after a negative result. By restricting the pathways for obtaining a positive result, gatekeeping controls the risk of false-positive results but preserves greater power for the earlier, higher-priority end points. This approach works well to test a sequence of secondary end points as in the CLEAN-TAVI study or to test a series of branching secondary end points (Figure).

Steps in serial gatekeeping are as follows: (1) determine the order for testing multiple end points, considering their relative importance and the likelihood that there is a difference in each; (2) test the first end point against the desired global false-positive rate (ie, .05) and, if the finding does not reach statistical significance, then stop all further testing and declare this and all downstream end points nonsignificant. If testing the first end point is significant, then declare this difference significant and proceed with the testing of the next end point; (3) test the next end point using a significance threshold of .05; if not significant, stop all further testing and declare this and all downstream end points nonsignificant. If significant, then declare this difference significant and proceed with the testing of the next end point; and (4) repeat the prior step until obtaining a first nonsignificant result, or until all end points have been tested.

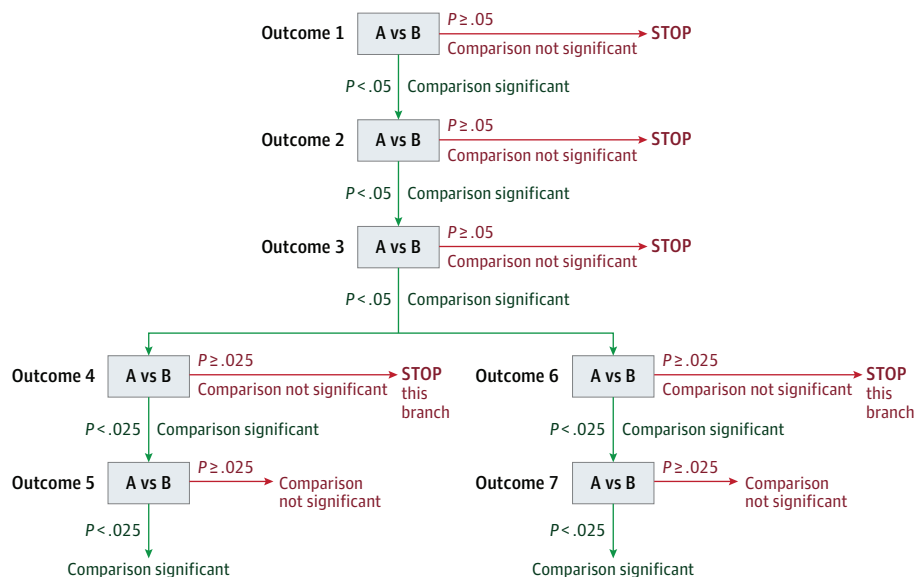
As shown in the Figure, this approach can be extended to test 2 or more end points at the same step by using a Bonferroni adjustment to evenly split the false-positive error rate within the step. In that case, testing is continued until either all branches have obtained a first nonsignificant result or all end points have been tested. For example, a neuroimaging end point could be used as a single end point for the first level, reflecting the assumption that if an improvement in an imaging outcome is not achieved then an improvement in a patient-centered functional outcome is highly unlikely, followed by a split to allow the testing of motor functions on one branch and verbal functions on the other. This avoids the need to prioritize either motor or verbal function over the other and may increase the ability to demonstrate an improvement in either domain.

Serial gatekeeping provides strict control of the false-positive error rate because it restricts multiple comparisons by sequentially testing hypotheses until the first nonsignificant test is found, and, *no matter how significant later end points appear to be*, they are never tested. The advantage is increased power for detecting effects on the end points that appear early in the sequence because they are tested against .05 rather than, eg, .05 divided by the total number of outcomes tested using a traditional Bonferroni adjustment. By accounting for the importance of certain hypotheses over others and by grouping hypotheses into primary and secondary groups, gatekeeping allocates the trial's power to be consistent with the investigators' priorities.⁶

What Are the Limitations of Gatekeeping Strategies?

Gatekeeping strategies are a powerful way to incorporate trial-specific clinical information to create prespecified ordering of hypotheses and mitigate the need to adjust for multiple comparisons

Figure. Criteria for Statistical Significance That Would Be Used in a Hypothetical Gatekeeping Strategy



This Figure shows the criteria for statistical significance that would be used in a hypothetical gatekeeping strategy in which there are 3 levels each with a single end point, followed by 2 levels with 2 end points each. The 3 end points are each tested in order against a criterion of .05. All testing stops as soon as 1 result is nonsignificant. If all are significant then a pair of fourth-level end points are tested, and to preserve the required significance of .05 at that level across 2

end points, the criterion for statistical significance is adjusted with a Bonferroni correction value of .025 for each. If 1 or both of these end points is significant at .025, then the next end point in the branch is tested, against a criterion of .025. If 1 or both are nonsignificant, no further testing occurs. If any outcome tested along a given pathway is not statistically significant, no further outcomes along that branch are tested because they are assumed to be nonsignificant.

at each stage of testing. The primary challenge in using gatekeeping is the need to prespecify and truly commit to the order of testing. The resulting limitation is that if, in retrospect, the order of outcome testing appears ill chosen (eg, if an early end point is negative and important end points later in the sequence appear to suggest large treatment effects), then there is no rigorous, post hoc method for statistically evaluating the later end points. This highlights the importance of having a clear data analysis strategy determined before the trial is started, and maintaining transparency (eg, publishing the study design and analysis plan on public websites or in journals).

How Was Gatekeeping Used in This Case?

The CLEAN-TAVI investigators used a gatekeeping strategy to compare several magnetic resonance imaging end points along with neurological and neurocognitive performance.¹ The first was the pri-

mary study end point, the number of brain lesions 2 days after TAVI. Secondary end points were only tested if the primary one was positive. Then, up to 16 secondary end points were tested in a defined sequence. The study was markedly positive, with the primary and many secondary end points demonstrating benefit. The first 8 comparisons were reported in detail in the publication—in their prespecified order—retaining the structure of the gatekeeping strategy.¹

How Should the Results Be Interpreted?

The CLEAN-TAVI clinical trial demonstrated the efficacy of a cerebral protection strategy with respect to multiple imaging measures of ischemic damage. The use of the prespecified gatekeeping strategy should provide assurance that the large number of imaging end points that were compared was unlikely to have led to false-positive results.

ARTICLE INFORMATION

Author Affiliations: Department of Emergency Medicine, Harbor-UCLA Medical Center, Torrance, California (Yadav, Lewis); Los Angeles Biomedical Research Institute, Torrance, California (Yadav); Berry Consultants, LLC, Austin, Texas (Lewis).

Corresponding Author: Kabir Yadav, MDCM, MS, MSHS, Department of Emergency Medicine, 1000 W Carson St, Box 21, Torrance, CA 90509 (kabir@emedharbor.edu).

Section Editors: Roger J. Lewis, MD, PhD, Department of Emergency Medicine, Harbor-UCLA Medical Center and David Geffen School of Medicine at UCLA; and Edward H. Livingston, MD, Deputy Editor, JAMA.

Conflict of Interest Disclosures: Both authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

REFERENCES

- Haussig S, Mangner N, Dwyer MG, et al. Effect of a cerebral protection device on brain lesions following transcatheter aortic valve implantation in patients with severe aortic stenosis. *JAMA*. 2016; 316(6):592-601.
- Cao J, Zhang S. Multiple comparison procedures. *JAMA*. 2014;312(5):543-544.

- Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. *BMJ*. 1995;310(6973):170-170.

- Hommel G, Bretz F, Maurer W. Powerful short-cuts for multiple testing procedures with special reference to gatekeeping strategies. *Stat Med*. 2007;26(22):4063-4073.

- Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat*. 1979;6(2):65-70.

- Dmitrienko A, Millen BA, Brechenmacher T, Paux G. Development of gatekeeping strategies in confirmatory clinical trials. *Biom J*. 2011;53(6):875-893.

Multiple Comparison Procedures

Jing Cao, PhD; Song Zhang, PhD

Problems can arise when researchers try to assess the statistical significance of more than 1 test in a study. In a single test, statistical significance is often determined based on an observed effect or finding that is unlikely (<5%) to occur due to chance alone. When more than 1 comparison is made, the chance of falsely detecting a non-existent effect increases. This is known as the problem of multiple comparisons (MCs), and adjustments can be made in statistical testing to account for this.¹

In this issue of *JAMA*, Saitz et al² report results of a randomized trial evaluating the efficacy of 2 brief counseling interventions (ie, a brief negotiated interview and an adaptation of a motivational interview, referred to as MOTIV) in reducing drug use in primary care patients when compared with not having an intervention. Because MCs were made, the authors adjusted how they determined statistical significance. In this article, we explain why adjustment for MCs is appropriate in this study and point out the limitations, interpretations, and cautions when using these adjustments.

Use of Method

Why Are Multiple Comparison Procedures Used?

When a single statistical test is performed at the 5% significance level, there is a 5% chance of falsely concluding that a supposed effect exists when in fact there is none. This is known as making a false discovery or having a false-positive inference. The significance level represents the risk of making a false discovery in an individual test, denoted as the individual error rate (IER). If 20 such tests are conducted, there is a 5% chance of making a false-positive inference with each test so that, on average, there will be 1 false discovery in the 20 tests.

Another way to view this is in terms of probabilities. If the probability of making a false conclusion (ie, false discovery) is 5% for a single test in which the effect does not exist, then 95% of the time, the test will arrive at the correct conclusion (ie, insignificant effect). With 2 such tests, the probability of finding an insignificant effect with the first test is 95%, as it is for the second. However, the probability of finding insignificant effects in the first and the second test is

0.95×0.95 , or 90%. With 20 such tests, the probability that all of the 20 tests correctly show insignificance is $(0.95)^{20}$ or 36%. So there is a 100% – 36%, or 64%, chance of at least 1 false-positive test occurring among the 20 tests. Because this probability quantifies the risk of making any false-positive inference by a group, or family, of tests, it is referred to as the family-wise error rate (FWER). The FWER generally increases as the number of tests performed increases. For example, assuming IER = 5% and denoting the number of multiple tests performed as K , then for $K = 2$ independent tests, $\text{FWER} = 1 - (0.95)^2 = 10\%$; for $K = 3$, $\text{FWER} = 1 - (0.95)^3 = 14\%$; and for $K = 20$, $\text{FWER} = 1 - (0.95)^{20} = 64\%$. This shows that the risk of making at least 1 false discovery in MCs can be greatly inflated even if the error rate is well controlled in each individual test.

When MCs are made, to control FWER at a certain level, the threshold for determining statistical significance in individual tests must be adjusted.¹ The simplest approach is known as the Bonferroni correction. It adjusts the statistical significance threshold by the number of tests. For example, for a FWER fixed at 5%, the IER in a group of 20 tests is set at $0.05/20 = 0.0025$; ie, an individual test would have to have a P value less than .0025 to be considered statistically significant. The Bonferroni correction is easy to implement, but it sets the significance threshold too rigidly, reducing the statistical procedure's power to detect true effects.

The Hochberg sequential procedure, which was used in the study by Saitz et al,² takes a different approach.³ All of the tests (the multiple comparisons) are performed and the resultant P values are ordered from largest to smallest on a list. If the FWER is fixed at 5% and the largest observed P value is less than .05, then all the tests are considered significant. Otherwise, if the next largest P value is less than $0.05/2$ (.025), then all the tests except the one with the largest P value are considered significant. If not, and the third P value in the list is less than $0.05/3$ (.017), then all the tests except those with the largest 2 P values are considered significant. This is continued until all the comparisons are made. This approach uses progressively more stringent statistical thresholds with the most stringent one being the Bonferroni threshold, and thus the approach can achieve a greater power to detect true effect than the Bonferroni procedure under appropriate conditions. An example in the Table consists of 6 tests in MCs; given a FWER of 5%, none of the tests

Table. An Example to Compare the Bonferroni Procedure and the Hochberg Sequential Procedure

Test	P Value	Bonferroni		Hochberg	
		Threshold	Result	Threshold	Result
1	.40	$0.05/6 = 0.008$	Not significant	0.05	Not significant
2	.027	$0.05/6 = 0.008$	Not significant	$0.05/2 = 0.025$	Not significant
3	.020	$0.05/6 = 0.008$	Not significant	$0.05/3 = 0.017$	Not significant
4	.012	$0.05/6 = 0.008$	Not significant	$0.05/4 = 0.0125$	Significant
5	.011	$0.05/6 = 0.008$	Not significant	NA	Significant
6	.010	$0.05/6 = 0.008$	Not significant	NA	Significant

Abbreviation: NA, not applicable.

are significant with the Bonferroni procedure. By comparison, 3 tests are significant with the Hochberg sequential procedure.

What Are the Limitations of Multiple Comparison Procedures?

Statistical procedures to control FWER in MCs were developed to reduce the risk of making any false-positive discovery. This is offset by having a lower test power to detect true effects. For example, when $K = 10$, the Bonferroni-corrected IER is $0.05/10 = 0.005$ to control FWER at 0.05. Under the conventional 2-sided t test, for a single test in the group to be considered significant, the observed effect needs to be 43% larger than that with an IER = 0.05. When $K = 20$, the Bonferroni-corrected IER is $0.05/20 = 0.0025$, and the observed effect needs to be 54% larger than that with an IER = 0.05. This limitation of reduced test power by controlling FWER becomes more apparent as the number of tests in MCs increases.

Why Did the Authors Use Multiple Comparison Procedures in This Particular Study?

In the study by Saitz et al, 2 tests were performed (brief negotiated interview vs no brief interview and MOTIV vs no brief interview) to determine if interventions with brief counseling were more effective in reducing drug use than interventions without counseling. With 2 tests and the IER set at 5%, the risk of falsely concluding at least 1 treatment is effective because of chance alone is 10%. To avoid the inflated FWER, the authors used the Hochberg sequential procedure.³

How Should This Method's Findings Be Interpreted in This Particular Study?

Saitz et al found that the adjusted P value⁴ based on the Hochberg procedure was .81 for both the brief negotiated interview and MOTIV vs no brief interview. The study did not provide sufficient evidence to claim that interventions with brief counseling were more effective than the one without brief counseling in reducing drug use among primary care patients. However, the absence of evidence does not mean there is an absence of an effect. The interventions may be effective, but this study did not have the statistical power to detect the effect.

Caveats to Consider When Looking at Multiple Comparison Procedures

To Adjust or Not

If researchers conduct multiple tests, each addressing an unrelated research question, then adjusting for MCs is unnecessary.

Suppose in a different study, brief negotiated interview was intended to treat alcohol use and MOTIV was intended to treat drug use. Then there is no need to adjust for MCs. This is in contrast to performing a family of tests from which the results as a whole address a single research question; then adjusting for MCs is necessary. As in the report by Saitz et al,² both the brief negotiated interview and MOTIV were compared with the control to draw a single conclusion about the efficacy of brief counseling interventions for drug use.

Confirmatory vs Exploratory

Bender and Lange⁵ suggested that MC procedures are only required for confirmatory studies for which the goal is the definitive proof of a predefined hypothesis to support final decision making. For exploratory studies seeking to generate hypotheses that will be tested in future confirmatory studies, the number of tests is usually large and the choice of hypotheses is likely data dependent (ie, selecting hypotheses after reviewing data), making MC adjustments unnecessary or even impossible at this stage of research. "Significant" results based on exploratory studies, however, should be clearly labeled so readers can correctly assess their scientific strength.

FWER vs FDR

The main approaches to MC adjustment include controlling FWER, which is the probability of making at least 1 false discovery in MCs, or controlling the false discovery rate (FDR), which is the expected proportion of false positives among all discoveries. When using the FDR approaches, a small proportion of false positives are tolerated to improve the chance of detecting true effects.⁶ In contrast, the FWER approaches avoid any false positives even at the cost of increased false negatives. The FDR and FWER represent 2 extremes of the relative importance of controlling for false positive or false negatives. The decision whether to control FWER or FDR should be made by carefully weighing the relative benefits between false-positive and false-negative discoveries in a specific study.

Definition of Family

Both FWER and FDR are defined for a particular family of tests. This "family" should be prespecified at the design stage of a study. Test bias can occur in MCs when selecting hypothesis to be tested after reviewing the data.

ARTICLE INFORMATION

Author Affiliations: Department of Statistical Science, Southern Methodist University, Dallas, Texas (Cao); Department of Clinical Sciences, UT Southwestern Medical Center, Dallas, Texas (Zhang).

Corresponding Author: Jing Cao, PhD, Southern Methodist University, Statistical Science, Dallas, TX 75205 (jcao@smu.edu).

Conflict of Interest Disclosures: Both authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

REFERENCES

- Hsu JC. *Multiple Comparisons: Theory and Methods*. London, UK: Chapman & Hall; 1996.
- Saitz R, Palfai TPA, Cheng DM, et al. Screening and brief intervention for drug use in primary care: the ASPIRE randomized clinical trial. *JAMA*. doi:10.1001/jama.2014.7862.
- Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*. 1988;75(4):800-802.
- Wright SP. Adjusted P value for simultaneous inference. *Biometrics*. 1992;48(4):1005-1013.
- Bender R, Lange S. Adjusting for multiple testing: when and how? *J Clin Epidemiol*. 2001;54(4):343-349.
- Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B*. 1995;57(1):289-300.

JAMA Guide to Statistics and Methods

Pragmatic Trials

Practical Answers to “Real World” Questions

Harold C. Sox, MD; Roger J. Lewis, MD, PhD

The concept of a “pragmatic” clinical trial was first proposed nearly 50 years ago as a study design philosophy that emphasizes answering questions of most interest to decision makers.¹ Decision makers,



Related article [page 1172](#)

ers, whether patients, physicians, or policy makers, need to know what they can expect from the available diagnostic or therapeutic options when applied in day-to-day clinical practice. This focus on addressing real-world effectiveness questions influences choices about trial design, patient population, interventions, outcomes, and analysis. In this issue of *JAMA*, Gottenberg et al² report the results of a trial designed to answer the question “If a biologic agent for rheumatoid arthritis is no longer effective for an individual patient, should the clinician recommend another drug with the same mechanism of action or switch to a biologic with a different mechanism of action?” Because the authors included some pragmatic elements in the trial design, this study illustrates the issues that clinicians should consider in deciding whether a trial result is likely to apply to their patients.

Use of the Method

Why Are Pragmatic Trials Conducted?

Pragmatic trials are intended to help typical clinicians and typical patients make difficult decisions in typical clinical care settings by maximizing the chance that the trial results will apply to patients that are usually seen in practice (external validity). The most important feature of a pragmatic trial is that patients, clinicians, clinical practices, and clinical settings are selected to maximize the applicability of the results to usual practice. Trial procedures and requirements must not inconvenience patients with substantial data collection and should impose a minimum of constraints on usual practice by allowing a choice of medication (within the constraints imposed by the purpose of the study) and dosage, providing the freedom to add cointerventions, and doing nothing to maximize adherence to the study protocol.

The pragmatic trial strategy contrasts with that used for an explanatory trial, the goal of which is to test a hypothesis that the intervention causes a clinical outcome. Explanatory trials seek to maximize the probability that the intervention—and not some other factor—causes the study outcome (internal validity). Explanatory trials seek to give the intervention the best possible chance to succeed by using experts to deliver it, delivering the intervention to patients who are most likely to respond, and administering the intervention in settings that provide expert after-care. Explanatory trials try to prevent any extraneous factors from influencing clinical outcomes, so they exclude patients who might have poor adherence and may intervene to maximize patient and clinician adherence to the study protocol. Explanatory trials are structured to

avoid downstream events that could affect study outcomes. If these events occur at different rates in the different study groups, the effect attributed to the intervention may be larger or smaller than its true effect. To avoid this problem, explanatory trials may choose a relatively short follow-up period. Explanatory trials pursue internal validity at the cost of external validity, whereas pragmatic trials place a premium on external validity while maintaining as much internal validity as possible.

Description of the Method

According to Tunis et al,³ “the characteristic features of [pragmatic clinical trials] are that they (1) select clinically relevant alternative interventions to compare, (2) include a diverse population of study participants, (3) recruit participants from heterogeneous practice settings, and (4) collect data on a broad range of health outcomes.” Eligible patients may be defined by presumptive diagnoses, rather than confirmed ones, because treatments are often initiated when the diagnosis is uncertain.³ Pragmatic trials may compare classes of drugs and allow the physician to choose which drug in the class to use, the dose, and any cointerventions, a freedom that mimics usual practice. Furthermore, the outcome measures are more likely to be patient-reported, global, subjective, and patient-centered (eg, self-reported quality-of-life measures), rather than the more disease-centered end points commonly used in explanatory trials (eg, the results of laboratory tests or imaging procedures).

Both approaches to study design must deal with the cost of clinical trials. Explanatory trials control costs by keeping the trial period as short as possible, consistent with the investigators’ ability to enroll enough patients to answer the study question. These trials preferentially recruit patients who will experience the study end point and not leave the study early because of disinterest or death from causes other than the target condition. Investigators in explanatory trials prefer to enroll participants with a high probability of experiencing an outcome in the near term. In contrast, pragmatic trials may control costs by leveraging existing data sources, eg, using disease registries to identify potential participants and using data in electronic health records to identify study outcomes.

Although these concepts sharpen the contrasts between pragmatic and explanatory trials for pedagogical reasons, in reality, many trials have features of both designs, in part to find a reasonable balance between internal validity and external validity.^{4,5}

What Are the Limitations of Pragmatic Trials?

The main limitation of a pragmatic trial is a direct consequence of choosing to conduct a lean study that puts few demands on patients and clinicians. Data collection may be sparse, and there are few clinical variables with which to identify subgroups of patients

who respond particularly well to one of the interventions. The use of the electronic health record as a source of data may save money, but it typically means inconsistent data collection and missing data. Relying on typical clinicians rather than experts in caring for patients with the target condition may lead to increased variability in practice and associated documentation of clinical findings. The variation caused by these shortcomings may reduce statistical precision and the capability of answering the research question unequivocally.

Why Was a Pragmatic Trial Conducted in This Case?

While Gottenber et al² cite the pragmatism of their study as its main strength, the authors do not explain their study design decisions. However, they imply a pragmatic motivation when they state that the study confirms the superiority of a drug from a different class in a setting that “corresponds to the therapeutic question clinicians face in daily practice.” The investigators note that their main limitation of the study was the inability to blind the participants to the identity of the drug they received. Blinding is especially important when the principal study outcomes are those reported by the patient, who may be influenced by knowing the intervention that they received.

How Should the Results Be Interpreted?

The study by Gottenberg et al² shows that, from the perspective of a population of patients, changing from one class of drugs to another improves the outcomes of care by rheumatologists in a rheumatology subspecialty clinic. This result has limited external validity. It probably applies to other rheumatology clinics, but its application to other settings is unknown. The main pragmatic feature of the study—allowing the rheumatologist to choose from several drugs within a class—implies that the main result applies strictly to the class of drugs rather than to any individual agent. It does not, for example, show that the improvement is the same regardless of

which within-class drug the clinician determines. The trial was also pragmatic in that clinicians were aware of the primary treatment and were free to choose cointerventions that would complement it, as would occur in clinical practice.

Several features of this study were not pragmatic, and others raise internal validity concerns. The researchers recruited participants from rheumatology specialty clinics. Although the article does not specify the clinicians who managed the patient's rheumatoid arthritis during the study, the clinicians were presumably rheumatologists in the participating practices. Even though the results apply to patients in a specialty clinic, whether they apply to patients managed by primary care physicians, with or without expert consultation, is unknown. The authors also did not specify the intensity of follow-up; was it typical of rheumatoid arthritis patients receiving biologic agents or did the study protocol specify more intensive follow up? The primary outcome measure was a score based on the erythrocyte sedimentation rate and a count of involved joints. The article does not identify the person who assessed the primary outcome. Assigning this task to the managing physician would be consistent with a pragmatic design, but it would also raise concerns about biased outcome assessment because the person measuring the outcome would know the treatment assignment.

The terms “explanatory” and “pragmatic” mark the ends of a spectrum of study designs. Typically, as noted by Thorpe and co-authors of the PRECIS (Pragmatic-Explanatory Continuum Indicator Summary) article,⁵ some features of a study are pragmatic and others are explanatory, as the study by Gottenberg et al illustrates and as would be expected because internal validity and external validity are typically achieved at the cost of one another. Whether the authors label their study as pragmatic or explanatory, readers should pay close attention to the study characteristics that maximize its applicability to their patients and their practice style.

ARTICLE INFORMATION

Author Affiliations: Patient Centered Outcomes Research Institute (PCORI), Washington, DC (Sox); Geisel School of Medicine at Dartmouth, Hanover, New Hampshire (Sox); Department of Emergency Medicine, Harbor-UCLA Medical Center, Torrance, California (Lewis); Department of Emergency Medicine, David Geffen School of Medicine at the University of California-Los Angeles (Lewis); Berry Consultants, LLC, Austin, Texas (Lewis).

Corresponding Author: Roger J. Lewis, MD, PhD, Department of Emergency Medicine, Harbor-UCLA Medical Center, 1000 W Carson St, PO Box 21, Torrance, CA 90509 (roger@emedharbor.edu).

Section Editors: Roger J. Lewis, MD, PhD, Department of Emergency Medicine, Harbor-UCLA Medical Center and David Geffen School of Medicine at UCLA; and Edward H. Livingston, MD, Deputy Editor, JAMA.

Conflict of Interest Disclosures: Both authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

Disclaimer: Dr Sox is an employee of the Patient-Centered Outcomes Research Institute (PCORI). This article does not represent the policies of PCORI.

REFERENCES

- Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutic trials. *J Chronic Dis*. 1967;20(8):637-648.
- Gottenberg J-E, Brocq O, Perdriger A, et al. Non-TNF-targeted biologic vs a second anti-TNF drug to treat rheumatoid arthritis in patients with insufficient response to a first anti-TNF drug. *JAMA*. doi:10.1001/jama.2016.13512
- Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. *JAMA*. 2003;290(12):1624-1632.
- Zwarenstein M, Treweek S, Gagnier JJ, et al; CONSORT group; Pragmatic Trials in Healthcare (Practihc) group. Improving the reporting of pragmatic trials: an extension of the CONSORT statement. *BMJ*. 2008;337:a2390. doi:10.1136/bmj.a2390
- Thorpe KE, Zwarenstein M, Oxman AD, et al. A Pragmatic-Explanatory Continuum Indicator Summary (PRECIS): a tool to help trial designers. *J Clin Epidemiol*. 2009;62(5):464-475.

JAMA Guide to Statistics and Methods

Equipoise in Research

Integrating Ethics and Science in Human Research

Alex John London, PhD

The principle of equipoise states that, when there is uncertainty or conflicting expert opinion about the relative merits of diagnostic, prevention, or treatment options, allocating interventions to individuals



Related article [page 483](#)

in a manner that allows the generation of new knowledge (eg, randomization) is ethically permissible.^{1,2} The principle of equipoise reconciles 2 potentially conflicting ethical imperatives: to ensure that research involving human participants generates scientifically sound and clinically relevant information while demonstrating proper respect and concern for the rights and interests of study participants.¹

In this issue of *JAMA*, Lascarrou et al³ report the results of a randomized trial designed to investigate whether the “routine use of the video laryngoscope for orotracheal intubation of patients in the ICU increased the frequency of successful first-pass intubation compared with use of the Macintosh direct laryngoscope.” Intubation in the intensive care unit (ICU) is associated with the potential for serious adverse events, and video laryngoscopy in the ICU has gained support from some clinicians who believe it to be superior to direct laryngoscopy. Such practitioners may therefore regard it as unethical to randomize study participants to direct laryngoscopy because they consider it to be an inferior intervention. But requiring uncertainty of individual clinicians to conduct a clinical trial gives too much ethical weight to personal judgment, hindering valuable research without providing benefit to patients. Therefore, it is important to understand the role of conflicting expert medical judgment in establishing equipoise and how this principle applies to the trial conducted by Lascarrou et al.

What Is Equipoise?

Two features of medical research pose special challenges for the goal of ensuring respect and concern for the rights and interests of participants. First, to generate reliable information, research often involves design features that alter the way participants are treated. For example, randomization and blinding are commonly used to reduce selection bias and treatment bias.⁴ Controlling how interventions are allocated and what researchers and participants know about who is receiving which interventions helps to more clearly distinguish the effects of the intervention from confounding effects. But randomization severs the link between what a participant receives and the recommendation of a treating clinician with an ethical duty to provide the best possible care for the individual person. In the study by Lascarrou et al,³ patients were randomized to undergo intubation with the video laryngoscope or the direct laryngoscope, independent of the preference of the treating physician.

Second, medical research involves exposing people to interventions whose risks and potential therapeutic, prophylactic, or diagnostic merits may be unknown, unclear, or the subject of disagree-

ment within the medical community. In the present case, some clinicians may maintain that video laryngoscopy is the superior strategy for orotracheal intubation in the ICU, others may disagree, while others judge that there is not sufficient evidence to warrant a strong commitment for or against this approach.

The principle of equipoise states that if there is uncertainty or conflicting expert opinion about the relative therapeutic, prophylactic, or diagnostic merits of a set of interventions, then it is permissible to allocate a participant to receive an intervention from this set, so long as there is not consensus that an alternative intervention would better advance that participant's interests.^{1,2,5-7}

In the present case, there is equipoise between video vs direct laryngoscopy because experts disagree about their relative clinical merits. These disagreements are reflected in variations in clinical practices. If it is ethically permissible for patients to receive care from expert clinicians in good professional standing with differing medical opinions about what constitutes optimal treatment, then it ordinarily cannot be wrong to permit participants to be randomized to those same treatment alternatives.⁵ Although randomization removes the link between what a participant receives and the recommendation of a particular clinician, the presence of equipoise ensures that each participant receives an intervention that would be recommended or utilized by at least a reasonable minority of informed expert clinicians.^{1,5,6} Equipoise thus ensures that randomization is consistent with respect for participant interests because it guarantees that no participant receives care known to be inferior to any available alternative.

Why Is Equipoise Important?

Ensuring equipoise helps researchers and institutional review boards (IRBs) fulfill 3 ethical obligations. First, to “disturb” equipoise studies must be designed to generate information that resolves uncertainty or reduces divergence in opinion among qualified medical experts. Such studies are likely to have both social and scientific value. Second, any risks to which participants are exposed must be reasonable in light of the value of the information a study is likely to produce.^{5,6} IRBs must make this determination before participants are enrolled.

Third is the obligation to show respect for potential participants as autonomous decision makers. Explaining during the informed consent process the nature of the uncertainty or conflict in medical judgment that a study is designed to resolve allows each individual to decide whether to participate by understanding the relevant uncertainties, their effects on that person's own interests, and how their resolution will contribute to improving the state of medical care.

What Are the Limitations of Equipoise?

Since its introduction, the concept of equipoise has received numerous formulations, creating the potential for confusion and

misunderstanding^{2,7} and spurring criticism and debate. One criticism holds that the version of equipoise described here is too permissive because it allows randomization even when individual clinicians are not uncertain about how best to treat a patient.⁸ The trial conducted by Lascarrou et al³ represents a case in which some clinicians have strong preferences for one modality of treatment over others. Requiring individual clinician uncertainty entrenches unwarranted variation in patient care by preventing participants from being offered the choice of participating in a study in which they might be allocated to interventions that would be recommended or utilized by other medical experts. If it is ethically acceptable for patients to receive care from informed, expert clinicians who favor different interventions, then it ordinarily cannot be unethical to allow patients to be randomized to the alternatives that such clinicians recommend. Legitimate disagreement among informed experts signifies that the clinical community lacks a basis for judging that patients are better off with one modality over the other.

An interpretation of equipoise that requires uncertainty on the part of the individual clinician is not ethically justified because it prevents studies that are likely to improve the quality of patient care without the credible expectation that this restriction will improve patient outcomes.

Another criticism is that equipoise is unlikely ever to exist, or to persist for long.⁹ This objection applies most directly to the view that equipoise only exists if the individual clinician believes that the interventions offered in a trial are of exactly equal expected value.¹⁰ On this view, equipoise would often disappear even though different experts retain conflicting medical recommendations. It therefore appears poorly suited to the goals of promoting the production of valuable information and protecting the interests of study participants.

How Is Equipoise Applied in This Case?

Lascarrou et al did not explicitly discuss equipoise in their study. However, the consent process approved by the ethics committee reflects the judgment that the interventions in the trial “were considered components of standard care” and patients who lacked decisional capacity could be enrolled even if no surrogate decision maker was present.

Ensuring that a study begins in and is designed to disturb a state of equipoise provides credible assurance to participants and other stakeholders that patients in medical distress can be enrolled in a study that will help improve patient care in emergency settings without concern that their health interests will be knowingly compromised in the process.

How Does Equipoise Influence the Interpretation of the Study?

In the past, strongly held beliefs about the effectiveness of treatments ranging from bloodletting to menopausal hormone therapy have proven to be false. Intubation in the ICU is associated with the potential for serious adverse events. Because video laryngoscopy is increasingly championed as the superior method for orotracheal intubation in the ICU, careful study of its relative merits and risks in comparison to conventional direct laryngoscopy addresses a question of clinical importance. The findings of Lascarrou et al³ suggest that perceived merits of video laryngoscopy do not translate into superior clinical outcomes and may be associated with higher rates of life-threatening complications. This result underscores the importance of conducting clinical research before novel interventions become widely incorporated into clinical practice, even if those interventions appear to offer clear advantages over existing alternatives.

ARTICLE INFORMATION

Author Affiliation: Department of Philosophy, Carnegie Mellon University, Pittsburgh, Pennsylvania.

Corresponding Author: Alex John London, PhD, Department of Philosophy, Center for Ethics and Policy, Carnegie Mellon University, 135 Baker Hall, Pittsburgh, PA 15213-3890 (ajlondon@andrew.cmu.edu).

Section Editors: Roger J. Lewis, MD, PhD, Department of Emergency Medicine, Harbor-UCLA Medical Center and David Geffen School of Medicine at UCLA; and Edward H. Livingston, MD, Deputy Editor, *JAMA*.

Conflict of Interest Disclosures: The author has completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

REFERENCES

1. Freedman B. Equipoise and the ethics of clinical research. *N Engl J Med*. 1987;317(3):141-145.
2. London AJ. Clinical equipoise: foundational requirement or fundamental error? In: Steinbock B, ed. *The Oxford Handbook of Bioethics*. Oxford, UK: Oxford University Press; 2007:571-595.
3. Lascarrou JB, Boisrame-Helms J, Bailly A, et al; Clinical Research in Intensive Care and Sepsis (CRICS) Group. Video laryngoscopy vs direct laryngoscopy on successful first-pass orotracheal intubation among ICU patients: a randomized clinical trial. *JAMA*. doi:10.1001/jama.2016.20603
4. Guyatt G, Rennie D, Meade MO, Cook DJ. *Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice*. 3rd ed. New York, NY: McGraw-Hill; 2015.
5. London AJ. Reasonable risks in clinical research: a critique and a proposal for the Integrative Approach. *Stat Med*. 2006;25(17):2869-2885.
6. Miller PB, Weijer C. Rehabilitating equipoise. *Kennedy Inst Ethics J*. 2003;13(2):93-118.
7. van der Graaf R, van Delden JJ. Equipoise should be amended, not abandoned. *Clin Trials*. 2011;8(4):408-416.
8. Hellman D. Evidence, belief, and action: the failure of equipoise to resolve the ethical tension in the randomized clinical trial. *J Law Med Ethics*. 2002;30(3):375-380.
9. Sackett DL. Equipoise, a term whose time (if it ever came) has surely gone. *CMAJ*. 2000;163(7):835-836.
10. Lilford RJ, Jackson J. Equipoise and the ethics of randomization. *J R Soc Med*. 1995;88(10):552-559.

JAMA Guide to Statistics and Methods

The Propensity Score

Jason S. Haukoos, MD, MSc; Roger J. Lewis, MD, PhD

Two recent studies published in *JAMA* involved the analysis of observational data to estimate the effect of a treatment on patient outcomes. In the study by Rozé et al,¹ a large observational data set was analyzed to estimate the relationship between early echocardiographic screening for patent ductus arteriosus and mortality among preterm infants. The authors compared mortality rates of 847 infants who were screened for patent ductus arteriosus and 666 who were not. The 2 infant groups were dissimilar; infants who were screened were younger, more likely female, and less likely to have received corticosteroids. The authors used propensity score matching to create 605 matched infant pairs from the original cohort to adjust for these differences. In the study by Huybrechts et al,² the Medicaid Analytic eXtract data set was analyzed to estimate the association between antidepressant use during pregnancy and persistent pulmonary hypertension of the newborn. The authors included 3 789 330 women, of which 128 950 had used antidepressants. Women who used antidepressants were different from those who had not, with differences in age, race/ethnicity, chronic illnesses, obesity, tobacco use, and health care use. The authors adjusted for these differences using, in part, the technique of propensity score stratification.

Use of the Method

Why Were Propensity Methods Used?

Many considerations influence the selection of one therapy over another. In many settings, more than one therapeutic approach is commonly used. In routine clinical practice, patients receiving one treatment will tend to be different from those receiving another, eg, if one treatment is thought to be better tolerated by elderly patients or more effective for patients who are more seriously ill. This results in a correlation—or confounding—between patient characteristics that affect outcomes and the choice of therapy (often called “confounding by indication”). If observational data obtained from routine clinical practice are examined to compare the outcomes of patients treated with different therapies, the observed difference will be the result of both differing patient characteristics and treatment choice, making it difficult to delineate the true effect of one treatment vs another.

The effect of an intervention is best assessed by randomizing treatment assignments so that, on average, the patients are similar in the 2 treatment groups. This allows a direct assessment of the effect of the intervention on outcome. In observational studies, randomization is not possible, so investigators must adjust for differences between groups to obtain valid estimates of the associations between the treatments being compared and the outcomes of interest.³ Multivariable statistical methods are often used to estimate this association while adjusting for confounding.

Propensity score methods are used to reduce the bias in estimating treatment effects and allow investigators to reduce the likelihood of confounding when analyzing nonrandomized, observational data. The propensity score is the probability that a patient would receive the treatment of interest, based on characteristics of the patient, treating

clinician, and clinical environment.⁴ Such probabilities can be estimated using multivariable statistical methods (eg, logistic regression), in which case the treatment of interest is the dependent variable and the characteristics of the patient, prescribing clinician, and clinical setting are the predictors. Investigators estimate these probabilities, ranging from 0 to 1, for each patient in the study population. These probabilities—the propensity scores—are then used to adjust for differences between groups. In biomedical studies, propensity scores are often used to compare treatments, but they can also be used to estimate the relationship between any nonrandomized factor, such as the exposure to a toxin or infectious agent and the outcome of interest.

There are 4 general ways propensity scores are used. The most common is *propensity score matching*, which involves assembling 2 groups of study participants, one group that received the treatment of interest and the other that did not, while matching individuals with similar or identical propensity scores.¹ The analysis of a propensity score–matched sample can then approximate that of a randomized trial by directly comparing outcomes between individuals who received the treatment of interest and those who did not, using methods that account for the paired nature of the data.⁵

The second approach is *stratification* on the propensity score.⁴ This technique involves separating study participants into distinct groups or strata based on their propensity scores. Five strata are commonly used, although increasing the number can reduce the likelihood of bias. The association between the treatment of interest and the outcome of interest is estimated within each stratum or pooled across strata to provide an overall estimate of the relationship between treatment and outcome. This technique relies on the notion that individuals within each stratum are more similar to each other than individuals in general; thus, their outcomes can be directly compared.

The third approach is *covariate adjustment* using the propensity score. For this approach, a separate multivariable model is developed, after the propensity score model, in which the study outcome serves as the dependent variable and the treatment group and propensity score serve as predictor variables. This allows the investigator to estimate the outcome associated with the treatment of interest while adjusting for the probability of receiving that treatment, thus reducing confounding.

The fourth approach is *inverse probability of treatment weighting* using the propensity score.⁶ In this instance, propensity scores are used to calculate statistical weights for each individual to create a sample in which the distribution of potential confounding factors is independent of exposure, allowing an unbiased estimate of the relationship between treatment and outcome.⁷

Alternative strategies—other than use of propensity scores—for adjusting for baseline differences between groups in observational studies include matching on baseline characteristics, performing stratified analyses, or using multivariable statistical methods to adjust for confounders. Propensity score methods are often more practical or statistically more efficient than these methods, in part

because propensity score methods can substantially limit the number of predictor variables used in the final analysis. Propensity score methods generally allow many more variables to be included in the propensity score model, which increases the ability of these approaches to effectively adjust for confounding, than could be incorporated directly into a multivariable analysis of the study outcome.

What Are the Limitations of Propensity Score Methods?

The propensity score for each study participant is based on the available measured patient characteristics, and unadjusted confounding may still exist if unmeasured factors influenced treatment selection. Therefore, using fewer variables in the propensity score model reduces the likelihood of effectively adjusting for confounding.

Although propensity score matching may be used to assemble comparable study groups, the quality of matching depends on the quality of the propensity score model, which in turn depends on the quality and size of the available data and how the model was built. Conventional modeling methods (eg, variable selection, use of interactions, regression diagnostics, etc) are not typically recommended for the development of propensity score models. For example, propensity score models may optimally include a larger number of predictor variables.

Why Did the Authors Use Propensity Methods?

In the reports by Rozé et al¹ and Huybrechts et al,² both of whom used propensity score methods because their data were observational, the treatments of interest (ie, screening by echocardiography and use of antidepressants in pregnancy) were not randomly allocated, and important characteristics differed between groups. Direct comparisons of the outcomes between treated and untreated groups would have likely resulted in significantly biased estimates. Instead, use of propensity score matching and stratification enabled the investigators to create study groups that were similar to one another and more accurately measure the relationship between treatment and outcome.

How Should the Findings Be Interpreted?

Given the observational nature of these studies, the fact that individuals in the treated and untreated groups were dissimilar, and the

goal of accurately estimating the association between treatment and outcome, the investigators had to adjust for differences in the treatment groups. Use of propensity score methods, whether by matching or stratification, resulted in less biased estimates than if such methods were not used. Even though observational data cannot definitely establish causal relationships or determine treatment effects as rigorously as a randomized clinical trial, assuming propensity score methods are properly used and the sample size is sufficiently large, these methods may provide a useful approximation of the likely effect of a treatment. This approach is particularly valuable for clinical situations in which randomized trials are not feasible or are unlikely to be conducted.

What Caveats Should the Reader Consider When Assessing the Results of Propensity Analyses?

The studies by Rozé et al¹ and Huybrechts et al² used propensity score matching and propensity score stratification, respectively. Although both methods are more valid in terms of balancing study groups than simple matching or stratification based on baseline characteristics, they vary in their ability to minimize bias. In general, propensity score matching minimizes bias to a greater extent than propensity score stratification. Assessment of balance between the groups, after use of propensity score methods, is important to allow readers to assess the comparability of patient groups.

Although no single standard approach exists to assess balance, comparing characteristics between treated and untreated patients typically begins with comparing summary statistics (eg, means or proportions) and the entire distributions of observed characteristics. For propensity score–matched samples, standardized differences (ie, differences divided by pooled standard deviations) are often used and, although no threshold is universally accepted, a standard difference less than 0.1 is often considered negligible. Assessing for balance provides a general sense for how well matching or stratification occurred and thus the extent to which the results are likely to be valid. Unfortunately, balance can only be demonstrated for patient characteristics that were measured in the study. Differences could still exist between patient groups that were not measured, resulting in biased results.

ARTICLE INFORMATION

Author Affiliations: Department of Emergency Medicine, University of Colorado School of Medicine, Denver (Haukoos); Department of Emergency Medicine, Harbor-UCLA Medical Center, Torrance, California (Lewis); David Geffen School of Medicine at UCLA, Los Angeles, California (Lewis).

Corresponding Author: Jason S. Haukoos, MD, MSc, Department of Emergency Medicine, Denver Health Medical Center, 777 Bannock St, Mail Code 0108, Denver, CO 80204 (Jason.Haukoos@dhha.org).

Section Editors: Roger J. Lewis, MD, PhD, Department of Emergency Medicine, Harbor-UCLA Medical Center and David Geffen School of Medicine at UCLA; and Edward H. Livingston, MD, Deputy Editor, *JAMA*.

Conflict of Interest Disclosures: All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

Funding/Support: Dr Haukoos is supported, in part, by grants R01AI106057 from the National Institute of Allergy and Infectious Diseases (NIAID) and R01HS021749 from the Agency for Healthcare Research and Quality (AHRQ).

Disclaimer: The views expressed herein are those of the authors and do not necessarily represent the views of NIAID, the National Institutes of Health, or AHRQ.

REFERENCES

1. Rozé JC, Cambonie G, Marchand-Martin L, et al; Hemodynamic EPIPAGE 2 Study Group. Association between early screening for patent ductus arteriosus and in-hospital mortality among extremely preterm infants. *JAMA*. 2015;313(24):2441-2448.
2. Huybrechts KF, Bateman BT, Palmsten K, et al. Antidepressant use late in pregnancy and risk of persistent pulmonary hypertension of the newborn. *JAMA*. 2015;313(21):2142-2151.
3. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology*. 1999;10(1):37-48.
4. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70:41-55.
5. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res*. 2011;46(3):399-424.
6. Schaffer JM, Singh SK, Reitz BA, Zamanian RT, Mallidi HR. Single- vs double-lung transplantation in patients with chronic obstructive pulmonary disease and idiopathic pulmonary fibrosis since the implementation of lung allocation based on medical need. *JAMA*. 2015;313(9):936-948.
7. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11(5):550-560.

Dose-Finding Trials

Optimizing Phase 2 Data in the Drug Development Process

Kert Viele, PhD; Jason T. Connor, PhD

Clinical trials in drug development are commonly divided into 3 categories or phases. The first phase aims to find the range of doses of potential clinical use, usually by identifying the maximum tolerated dose. The second phase



Related article [page 2251](#)

aims to find doses that demonstrate promising efficacy with acceptable safety. The third phase aims to confirm the benefit previously found in the second phase using clinically meaningful end points and to demonstrate safety more definitively.

Dose-finding trials—studies conducted to identify the most promising doses or doses to use in later studies—are a key part of the second phase and are intended to answer the dual questions of whether future development is warranted and what dose or doses should be used. If too high a dose is chosen, adverse effects in later confirmatory phase 3 trials may threaten the development program. If too low a dose is chosen, the treatment effect may be too small to yield a positive confirmatory trial and gain approval from a regulatory agency. A well-designed dose-finding trial is able to establish the optimal dose of a medication and facilitate the decision to proceed with a phase 3 trial.

Selection of a dose for further testing requires an understanding of the relationships between dose and both efficacy and safety. These relationships can be assessed by comparing the data from each dose group with placebo, or with the other doses, in a series of pairwise comparisons. This approach is prone to both false-negative and false-positive results because of the large number of statistical comparisons and the relatively small number of patients receiving each dose. These risks can be mitigated by combining data from patients receiving multiple active doses into a single treatment group for comparison with placebo (“pooling”), but only if it is possible to reliably predict which doses are likely to be effective.

In general, dose-response relationships are best examined through dose-response models that make flexible, justifiable assumptions about the potential dose-response relationships and allow the integration of information from all doses used in the trial. This can reduce the risk of both false-negative and false-positive results; incorporating all data into the estimates of efficacy and safety for each dose produces more accurate estimates than evaluating the response to each dose separately.

In this issue of *JAMA*, Gheorghiade et al¹ report the results of SOCRATES-REDUCED, a randomized placebo-controlled dose-finding clinical trial investigating 4 different target doses of vericiguat for patients with worsening chronic heart failure, with the primary outcome being a reduction in log-transformed level of N-terminal pro-B-type natriuretic peptide. The primary approach to analyzing the dose response, combining the data from patients allocated to the 3 highest target doses (pooling) for comparison with placebo,

yielded a negative result ($P = .15$), but a different dose-response model based on linear regression, used in an exploratory secondary analysis, yielded a positive result ($P = .02$).

Use of the Method

Why Are Dose-Response Models Used?

A dose-response model assumes a general relationship between dose and efficacy or dose and the rates of adverse effects.² Ideally, this allows data from patients receiving all doses of the drug to contribute to the estimated dose-response curve, maximizing the statistical power of the study and reducing the uncertainty in the estimates of the effects of each dose. When a sufficiently flexible general relationship is used, the dose-response model correctly identifies doses of low or high efficacy (avoiding the assumption of similar efficacy across doses, as is implied with pooling) while smoothing out spurious highs and lows (avoiding problems that occur when each dose is analyzed separately). A model can produce estimates and confidence intervals for the effect of every dose and often even for drug doses not included in the trial.

Dose-response modeling is first used to determine whether a treatment effect appears to exist and, if so, to estimate dose-specific effects to help optimize subsequent phase 3 trial design. Unlike a confirmatory trial in which a regulatory agency makes a binary decision (eg, to approve or not approve a drug), phase 2 trials are used to inform the next stage of drug development. Therefore, estimation of the magnitude of treatment effects is more important than testing hypotheses regarding treatment effects. Phase 2 dose-finding studies can also be used to predict the likelihood of later phase 3 success through calculation of predictive probabilities.³

The assumptions in the dose-response model can be rigid or flexible to match preexisting knowledge of the clinical setting. When accurate, such assumptions can increase the power of a trial design by incorporating known clinical information. When inaccurate, these assumptions compromise the statistical properties of the trial and the interpretability of the results. For example, in SOCRATES-REDUCED, the primary analysis consisted of pooling data from the 3 highest-dose regimens.¹ This approach is most effective when the efficacious region of the dose range can be predicted reliably. The exploratory secondary analysis in SOCRATES-REDUCED was based on a linear regression model. This approach is most effective when a linear dose-response relationship is likely to exist over the range of doses evaluated in the trial.

A common dose-response model is the E_{max} model,⁴ which assumes an S-shaped curve for the dose response (eg, a monotonically increasing curve that is flat for low doses, increases for the middle dose range, and then flattens out again for high doses). The model is flexible in that the height of the plateau, the dose location of the increase in efficacy, and the rate of increase may all be in-

formed by the data. Alternatives to the E_{\max} model include smoothing models such as a normal dynamic linear model.⁵ These models take the raw data and produce a smooth curve that eliminates random highs and lows but maintains the general shape. Normal dynamic linear models are particularly useful for dose responses that may be “inverted U” shaped and may be applicable when the dose response is for an outcome that combines safety and efficacy (low doses may not be efficacious, high doses may be unsafe, resulting in an inverted U shape, with the optimal dose in the middle).

What Are the Limitations of Dose-Response Modeling?

All dose-response models require assumptions regarding the potential shapes of the dose-response curve, although sometimes (eg, with pooling) the assumptions are only implied. When assumptions are incorrect, inferences from the model may be invalid. In SOCRATES-REDUCED, the implied assumption of the primary analysis of similar efficacy among the 3 highest doses was not supported by the data. Similarly, the linear model used in the exploratory secondary analysis assumed that the increase in benefit from one dose to the next was the same between every successive pair of doses. This also does not appear to be strictly consistent with the data obtained in the trial.

Why Did the Authors Use Dose-Response Modeling in This Particular Study?

The authors used dose-response modeling to maximize the power of the primary analysis hypothesis test. If the 3 highest doses had all been similarly effective, pooling of data from these doses would result in higher sample sizes in the treatment group of the primary

“treatment vs placebo” hypothesis test and higher power to detect an effect. In the exploratory secondary analysis using the linear dose-response model, the authors used a model that allowed the higher doses to be significantly more efficacious than the lower doses.

How Should the Dose-Response Findings Be Interpreted in This Particular Study?

Figure 2 in the report by Gheorghiadu et al¹ shows the key dose-response relationship and suggests that the 10-mg target dose is the most or possibly only effective dose. However, the primary analysis was null, and the protocol called for the statistical secondary analysis only if the primary analysis were significant at $P < .05$. Therefore, although the 10-mg dose appears to be the most promising for investigation in a phase 3 trial, the dose-ranging findings must be considered very tentative. There remains uncertainty regarding how best to estimate the effect of the 10-mg dose. The primary analysis did not evaluate the effect of the 10-mg dose alone, and separate analyses for each dose would be prone to high variation and false-positive results due to multiple comparisons. The exploratory linear model produced an estimated effect for the 10-mg dose under an assumption of linearity. This analysis and its results were considered only exploratory.

Caveats to Consider When Looking at Results Based on a Dose-Response Model

It is often useful to inspect a plot of the dose-response model-based estimates against all data observed in the trial. This allows visual confirmation that the chosen dose-response model captures the general shape of the observed data.

ARTICLE INFORMATION

Author Affiliations: Berry Consultants LLC, Austin, Texas (Viele, Connor); University of Central Florida College of Medicine, Orlando (Connor).

Corresponding Author: Kert Viele, PhD, Berry Consultants LLC, 4301 Westbank Dr, Bldg B, Ste 140, Austin, TX 78746 (kert@berryconsultants.com).

Section Editors: Roger J. Lewis, MD, PhD, Department of Emergency Medicine, Harbor-UCLA Medical Center and David Geffen School of Medicine at UCLA; and Edward H. Livingston, MD, Deputy Editor, *JAMA*.

Conflict of Interest Disclosures: Both authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

REFERENCES

1. Gheorghiadu M, Greene SJ, Butler J, et al. Effect of vericiguat, a soluble guanylate cyclase stimulator, on natriuretic peptide levels in patients with worsening chronic heart failure and reduced ejection fraction: the SOCRATES-REDUCED randomized trial. *JAMA*. doi:10.1001/jama.2015.15734.
2. Bretz F, Hsu J, Pinheiro J, Liu Y. Dose finding: a challenge in statistics. *Biom J*. 2008;50(4):480-504.
3. Saville BR, Connor JT, Ayers GD, Alvarez J. The utility of Bayesian predictive probabilities for interim monitoring of clinical trials. *Clin Trials*. 2014;11(4):485-493.
4. Dragalin V, Hsuan F, Padmanabhan SK. Adaptive designs for dose-finding studies based on sigmoid Emax model. *J Biopharm Stat*. 2007;17(6):1051-1070.
5. Krams M, Lees KR, Hacke W, Grieve AP, Orgogozo JM, Ford GA; ASTIN Study Investigators. Acute Stroke Therapy by Inhibition of Neutrophils (ASTIN): an adaptive dose-response study of UK-279,276 in acute ischemic stroke. *Stroke*. 2003;34(11):2543-2548.

JAMA Guide to Statistics and Methods

Odds Ratios—Current Best Practice and Use

Edward C. Norton, PhD; Bryan E. Dowd, PhD; Matthew L. Maciejewski, PhD

Odds ratios frequently are used to present strength of association between risk factors and outcomes in the clinical literature. Odds and odds ratios are related to the probability of a binary outcome (an outcome that is either present or absent, such as mortality). The *odds* are the ratio of the probability that an outcome occurs to the probability that the outcome does not occur. For example, suppose that the probability of mortality is 0.3 in a group of patients. This can be expressed as the odds of dying: $0.3/(1 - 0.3) = 0.43$. When the probability is small, odds are virtually identical to the probability. For example, for a probability of 0.05, the odds are $0.05/(1 - 0.05) = 0.052$. This similarity does not exist when the value of a probability is large.

Probability and odds are different ways of expressing similar concepts. For example, when randomly selecting a card from a deck, the probability of selecting a spade is $13/52 = 25\%$. The odds of selecting a card with a spade are $25\%/75\% = 1:3$. Clinicians usually are interested in knowing probabilities, whereas gamblers think in terms of odds. Odds are useful when wagering because they represent fair payouts. If one were to bet \$1 on selecting a spade from a deck of cards, a payout of \$3 is necessary to have an even chance of winning your money back. From the gambler's perspective, a payout smaller than \$3 is unfavorable and greater than \$3 is favorable.

Differences between 2 different groups having a binary outcome such as mortality can be compared using odds ratios, the ratio of 2 odds. Differences also can be compared using probabilities by calculating the *relative risk ratio*, which is the ratio of 2 probabilities. Odds ratios commonly are used to express strength of associations from logistic regression to predict a binary outcome.¹

Why Report Odds Ratios From Logistic Regression?

Researchers often analyze a binary outcome using multivariable logistic regression. One potential limitation of logistic regression is that the results are not directly interpretable as either probabilities or relative risk ratios. However, the results from a logistic regression are converted easily into odds ratios because logistic regression estimates a parameter, known as the log odds, which is the natural logarithm of the odds ratio. For example, if a log odds estimated by logistic regression is 0.4 then the odds ratio can be derived by exponentiating the log odds ($\exp(0.4) = 1.5$). It is the odds ratio that is usually reported in the medical literature. The odds ratio is always positive, although the estimated log odds can be positive or negative (log odds of -0.2 equals odds ratio of $0.82 = \exp(-0.2)$).

The odds ratio for a risk factor contributing to a clinical outcome can be interpreted as whether someone with the risk factor is more or less likely than someone without that risk factor to experience the outcome of interest. Logistic regression modeling allows the estimates for a risk factor of interest to be adjusted for other risk factors, such as age, smoking status, and diabetes. One nice feature of the logistic function is that an odds ratio for one covariate is constant for all values of the other covariates.

Another nice feature of odds ratios from a logistic regression is that it is easy to test the statistical strength of association. The stan-

dard test is whether the parameter (log odds) equals 0, which corresponds to a test of whether the odds ratio equals 1. Odds ratios typically are reported in a table with 95% CIs. If the 95% CI for an odds ratio does not include 1.0, then the odds ratio is considered to be statistically significant at the 5% level.

What Are the Limitations of Odds Ratios?

Several caveats must be considered when reporting results with odds ratios. First, the interpretation of odds ratios is framed in terms of odds, not in terms of probabilities. Odds ratios often are mistaken for relative risk ratios.^{2,3} Although for rare outcomes odds ratios approximate relative risk ratios, when the outcomes are not rare, odds ratios always overestimate relative risk ratios, a problem that becomes more acute as the baseline prevalence of the outcome exceeds 10%. Odds ratios cannot be calculated directly from relative risk ratios. For example, an odds ratio for men of 2.0 could correspond to the situation in which the probability for some event is 1% for men and 0.5% for women. An odds ratio of 2.0 also could correspond to a probability of an event occurring 50% for men and 33% for women, or to a probability of 80% for men and 67% for women.

Second, and less well known, the magnitude of the odds ratio from a logistic regression is scaled by an arbitrary factor (equal to the square root of the variance of the unexplained part of binary outcome).⁴ This arbitrary scaling factor changes when more or better explanatory variables are added to the logistic regression model because the added variables explain more of the total variation and reduce the unexplained variance. Therefore, adding more independent explanatory variables to the model will increase the odds ratio of the variable of interest (eg, treatment) due to dividing by a smaller scaling factor. In addition, the odds ratio also will change if the additional variables are not independent, but instead are correlated with the variable of interest; it is even possible for the odds ratio to decrease if the correlation is strong enough to outweigh the change due to the scaling factor.

Consequently, there is no unique odds ratio to be estimated, even from a single study. Different odds ratios from the same study cannot be compared when the statistical models that result in odds ratio estimates have different explanatory variables because each model has a different arbitrary scaling factor.⁴⁻⁶ Nor can the magnitude of the odds ratio from one study be compared with the magnitude of the odds ratio from another study, because different samples and different model specifications will have different arbitrary scaling factors. A further implication is that the magnitudes of odds ratios of a given association in multiple studies cannot be synthesized in a meta-analysis.⁴

How Did the Authors Use Odds Ratios?

In a recent *JAMA* article, Tringale and colleagues⁷ studied industry payments to physicians for consulting, ownership, royalties, and research as well as whether payments differed by physician specialty or sex. Industry payments were received by 50.8% of men across

all specialties compared with 42.6% of women across all specialties. Converting these probabilities to odds, the odds that men receive industry payments is 1.03 (0.51/0.49), and the odds that women receive industry payments is 0.74 = (0.43/0.57).

The odds ratio for men compared with women is the ratio of the odds for men divided by the odds for women. In this case, the unadjusted odds ratio is $1.03/0.74 = 1.39$. Therefore, the odds for men receiving industry payments are about 1.4 as large (40% higher) compared with women. Note that the ratio of the odds is different than the ratio of the probabilities because the probability is not close to 0. The unadjusted ratio of the probabilities for men and women (Tringale et al⁷ report each probability, but not the ratio), the relative risk ratio, is 1.19 (0.51/0.43).

Greater odds that men may receive industry payments may be explained by their disproportionate representation in specialties more likely to receive industry payments. After controlling for specialty (and other factors), the estimated odds ratio was reduced from 1.39 to 1.28, with a 95% CI of 1.26 to 1.31, which did not include 1.0 and, therefore, is statistically significant. The odds ratio probably declined after adjusting for more variables because they were correlated with physicians' sex.

How Should the Findings Be Interpreted?

In exploring the association between physician sex and receiving industry payments, Tringale and colleagues⁷ found that men are

more likely to receive payments than women, even after controlling for confounders. The magnitude of the odds ratio, about 1.4, indicates the direction of the effect, but the magnitude of the number itself is hard to interpret. The estimated odds ratio is 1.4 when simultaneously accounting for specialty, spending region, sole proprietor status, sex, and the interaction between specialty and sex. A different odds ratio would be found if the model included a different set of explanatory variables. The 1.4 estimated odds ratio should not be compared with odds ratios estimated from other data sets with the same set of explanatory variables, or to odds ratios estimated from this same data set with a different set of explanatory variables.⁴

What Caveats Should the Reader Consider?

Odds ratios are one way, but not the only way, to present an association when the main outcome is binary. Tringale et al⁷ also report absolute rate differences. The reader should understand odds ratios in the context of other information, such as the underlying probability. When the probabilities are small, odds ratios and relative risk ratios are nearly identical, but they can diverge widely for large probabilities. The magnitude of the odds ratio is hard to interpret because of the arbitrary scaling factor and cannot be compared with odds ratios from other studies. It is best to examine study results presented in several ways to better understand the true meaning of study findings.

ARTICLE INFORMATION

Author Affiliations: Department of Health Management and Policy, Department of Economics, University of Michigan, Ann Arbor (Norton); National Bureau of Economic Research, Cambridge, Massachusetts (Norton); Division of Health Policy and Management, School of Public Health, University of Minnesota, Minneapolis (Dowd); Center for Health Services Research in Primary Care, Durham Veterans Affairs Medical Center, Durham, North Carolina (Maciejewski); Department of Population Health Sciences, Duke University School of Medicine, Durham, North Carolina (Maciejewski); Division of General Internal Medicine, Department of Medicine, Duke University School of Medicine, Durham, North Carolina (Maciejewski).

Section Editors: Roger J. Lewis, MD, PhD, Department of Emergency Medicine, Harbor-UCLA Medical Center and David Geffen School of Medicine at UCLA; and Edward H. Livingston, MD, Deputy Editor, *JAMA*.

Conflict of Interest Disclosures: All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Dr Maciejewski reported receiving personal fees from the University of Alabama at Birmingham for a workshop presentation; receiving grants from NIDA and the Veterans Affairs; receiving a contract from NCQA to Duke University for research; being supported by a research career scientist award 10-391 from the Veterans Affairs Health Services Research and Development; and that his spouse owns stock in Amgen. No other disclosures were reported.

REFERENCES

1. Meurer WJ, Tolles J. Logistic regression diagnostics: understanding how well a model predicts outcomes. *JAMA*. 2017;317(10):1068-1069. doi:10.1001/jama.2016.20441
2. Schwartz LM, Woloshin S, Welch HG. Misunderstandings about the effects of race and sex on physicians' referrals for cardiac

catheterization. *N Engl J Med*. 1999;341(4):279-283. doi:10.1056/NEJM199907223410411

3. Holcomb WL Jr, Chaiworapongsa T, Luke DA, Burgdorf KD. An odd measure of risk: use and misuse of the odds ratio. *Obstet Gynecol*. 2001;98(4):685-688.

4. Norton EC, Dowd BE. Log odds and the interpretation of logit models. *Health Serv Res*. 2018;53(2):859-878. doi:10.1111/1475-6773.12712

5. Miettinen OS, Cook EF. Confounding: essence and detection. *Am J Epidemiol*. 1981;114(4):593-603. doi:10.1093/oxfordjournals.aje.a113225

6. Hauck WW, Neuhaus JM, Kalbfleisch JD, Anderson S. A consequence of omitted covariates when estimating odds ratios. *J Clin Epidemiol*. 1991;44(1):77-81. doi:10.1016/0895-4356(91)90203-L

7. Tringale KR, Marshall D, Mackey TK, Connor M, Murphy JD, Hattangadi-Gluth JA. Types and distribution of payments from industry to physicians in 2015. *JAMA*. 2017;317(17):1774-1784. doi:10.1001/jama.2017.3091

JAMA Guide to Statistics and Methods

Evaluating Discrimination of Risk Prediction Models

The C Statistic

Michael J. Pencina, PhD; Ralph B. D'Agostino Sr, PhD

Risk prediction models help clinicians develop personalized treatments for patients. The models generally use variables measured at one time point to estimate the probability of an outcome occurring within a given time in the future. It is essential to assess the performance of a risk prediction model in the setting in which it will be used. This is done by evaluating the model's discrimination and calibration. *Discrimination* refers to the ability of the model to separate individuals who develop events from those who do not. In time-to-event settings, discrimination is the ability of the model to predict who will develop an event earlier and who will develop an event later or not at all. *Calibration* measures how accurately the model's predictions match overall observed event rates.

In this issue of *JAMA*, Melgaard et al used the C statistic, a global measure of model discrimination, to assess the ability of the CHA₂DS₂-VASC model to predict ischemic stroke, thromboembolism, or death in patients with heart failure and to do so separately for patients who had or did not have atrial fibrillation (AF).¹

Use of the Method

Why Are C Statistics Used?

The C statistic is the probability that, given 2 individuals (one who experiences the outcome of interest and the other who does not or who experiences it later), the model will yield a higher risk for the first patient than for the second. It is a measure of concordance (hence, the name "C statistic") between model-based risk estimates and observed events. C statistics measure the ability of a model to rank patients from high to low risk but do not assess the ability of a model to assign accurate probabilities of an event occurring (that is measured by the model's calibration). C statistics generally range from 0.5 (random concordance) to 1 (perfect concordance).

C statistics can also be thought of as being the area under the plot of sensitivity (proportion of people with events for whom the model predicts are high risk) vs 1 minus specificity (proportion of people without events for whom the model predicts are high risk) for all possible classification thresholds. This plot is called the receiver operating characteristic (ROC) curve, and the C statistic is equal to the area under this curve.² For example, in the study by Melgaard et al, CHA₂DS₂-VASC scores ranged from a low of 0 (heart failure only) to a high of 5 or higher, depending on the number of comorbidities a patient had. One point on the ROC curve would be when high risk is defined as a CHA₂DS₂-VASC score of 1 or higher and low risk as a CHA₂DS₂-VASC score of 0. Another point on the curve would be when high risk is defined as a CHA₂DS₂-VASC score of 2 or higher and low risk as a CHA₂DS₂-VASC score of lower than 2, etc. Each cut point is associated with a different sensitivity and specificity.

It is useful to quantify the performance and clinical value of predictive models using the positive predictive value (PPV; the proportion of patients in whom the model predicts an event will occur who actually have an event) and the negative predictive value (NPV; the proportion of patients whom the model predicts will not have an event who actually do not experience the event). An important measure of a model's misclassification of events is 1 minus NPV, or the proportion of patients the model predicts will not have an event who actually have the event. The PPV and 1 minus NPV can be more informative for individual patients than the sensitivity and specificity because they answer the question "What are this patient's chances of having an event when the model predicts they will or will not have one?" If the event rate is known, then the PPV and NPV can be estimated based on sensitivity and specificity and, hence, the C statistic can be viewed as a summary for both sets of measures.

What Are the Limitations of the C Statistic?

The C statistic has several limitations. As a single number, it summarizes the discrimination of a model but does not communicate all the information ROC plots contain and lacks direct clinical application. The NPV, PPV, sensitivity, and specificity have more clinical relevance, especially when presented as plots across all meaningful classification thresholds (as is done with ROCs). A weighted sum of sensitivity and specificity (known as the standardized net benefit) can be plotted to assign different penalties to the 2 misclassification errors (predicting an individual who ultimately experiences an event to be at low risk; predicting an individual who does not experience an event to be at high risk) according to the principles of decision analysis.^{3,4} In contrast, the C statistic does not effectively balance misclassification errors.⁵ In addition, the C statistic is only a measure of discrimination, not calibration, so it provides no information regarding whether the overall magnitude of risk is predicted accurately.

Why Did the Authors Use C Statistics in Their Study?

Melgaard et al¹ sought to determine if the CHA₂DS₂-VASC score could predict occurrences of ischemic stroke, thromboembolism, or death among patients who have heart failure with and without AF. The authors used the C statistic to determine how well the model could distinguish between patients who would or would not develop each of the 3 end points they studied. The C statistic yielded the probability that a randomly selected patient who had an event had a risk score that was higher than a randomly selected patient who did not have an event.

How Should the Findings Be Interpreted?

The value of the C statistic depends not only on the model under investigation (ie, CHA₂DS₂-VASC score) but also on the distribution of risk factors in the sample to which it is applied. For example, if age is an important risk factor, the same model can appear to perform

much better when applied to a sample with a wide age range compared with a sample with a narrow age range.

The C statistics reported by Melgaard et al¹ range from 0.62 to 0.71 and do not appear impressive (considering that a C statistic of 0.5 represents random concordance). This might be due to limitations of the model; eg, if there were an insufficient number of predictors or the predictors had been dichotomized for simplicity. The nationwide nature of the data used by Melgaard et al suggests that the unimpressive values of the C statistic cannot be attributed to narrow ranges of risk factors in the analyzed cohort. Rather, it might suggest inherent limitations in the ability to discriminate between patients with heart failure who will and will not die or develop ischemic stroke or thromboembolism.

The C statistic analysis suggested that the CHA₂DS₂-VASc model performed similarly among heart failure patients with and without AF (C statistics between 0.62 and 0.71 among patients with AF and 0.63 to 0.69 among patients without AF). An additional insight emerges from NPV analysis looking at misclassification of events occurring at 5 years, however. Between 19% and 27% of patients without AF who were predicted to be at low risk actually had 1 of the 3 events and thus were misclassified, yielding an NPV of 73% to 82%. Between 24% and 39% of patients with AF whom the model clas-

sified as low risk had major events, yielding an NPV of 61% to 76%. Because there was less misclassification among patients without AF who were predicted to be at low risk, a CHA₂DS₂-VASc score of 0 is a better determinant of long-term low risk among patients without AF than patients with AF. This aspect of the model performance is not apparent when looking at C statistics alone.

Caveats to Consider When Using C Statistics to Assess Predictive Model Performance

Special extensions of the C statistic need to be used when applying it to time-to-event data⁶ and competing-risk settings.⁷ Furthermore, there exist several appealing single-number alternatives to the C statistic. They include the discrimination slope, the Brier score, or the difference between sensitivity and 1 minus specificity evaluated at the event rate.³

The C statistic provides an important but limited assessment of the performance of a predictive model and is most useful as a familiar first-glance summary. The evaluation of the discriminating value of a risk model should be supplemented with other statistical and clinical measures. Graphical summaries of model calibration and clinical consequences of adopted decisions are particularly useful.⁸

ARTICLE INFORMATION

Author Affiliations: Biostatistics and Bioinformatics, Duke Clinical Research Institute, Duke University, Durham, North Carolina (Pencina); Department of Mathematics and Statistics, Boston University, Boston, Massachusetts (D'Agostino).

Corresponding Author: Michael J. Pencina, PhD, Biostatistics and Bioinformatics, Duke Clinical Research Institute, Duke University, 2400 Pratt St, Durham, NC 27705 (michael.pencina@duke.edu).

Section Editors: Roger J. Lewis, MD, PhD, Department of Emergency Medicine, Harbor-UCLA Medical Center and David Geffen School of Medicine at UCLA; and Edward H. Livingston, MD, Deputy Editor, *JAMA*.

Conflict of Interest Disclosures: The authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

REFERENCES

- Melgaard L, Gorst-Rasmussen A, Lane DA, Rasmussen LH, Larsen TB, Lip GYH. Assessment of the CHA₂DS₂-VASc score in predicting ischemic stroke, thromboembolism, and death in patients with heart failure with and without atrial fibrillation. *JAMA*. doi:10.1001/jama.2015.10725.
- Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29-36.
- Pepe MS, Janes H. Methods for evaluating prediction performance of biomarkers and tests. In: Lee M-LT, Gail M, Pfeiffer R, Satten G, Cai T, Gandy A, eds. *Risk Assessment and Evaluation of Predictions*. New York, NY: Springer; 2013:107-142.
- Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26(6):565-574.
- Hand DJ. Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Mach Learn*. 2009;77:103-123.
- Pencina MJ, D'Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Stat Med*. 2004;23(13):2109-2123.
- Blanche P, Dartigues JF, Jacqmin-Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Stat Med*. 2013;32(30):5381-5397.
- Moons KGM, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med*. 2015;162(1):W1-W73.

JAMA Guide to Statistics and Methods

Time-to-Event Analysis

Juliana Tolles, MD, MHS; Roger J. Lewis, MD, PhD

Time-to-event analysis, also called survival analysis, was used in the study by Nissen et al¹ published in this issue of *JAMA* to compare the risk of major adverse cardiovascular events (MACE) in a noninferiority trial of a combination of naltrexone and bupropion vs placebo for overweight or obese patients with cardiovascular risk factors. The authors used a type of time-to-event analysis called Cox proportional hazards modeling to compare the risk of MACE in the 2 groups, concluding that the use of naltrexone-bupropion increased the risk of MACE per unit time by no more than a factor of 2.



Related article [page 990](#)

Use of the Method

Why Is Time-to-Event Analysis Used?

One way to evaluate how a medical treatment affects patients' risk of an adverse outcome is to analyze the time intervals between the initiation of treatment and the occurrence of such events. That information can be used to calculate the hazard for each treatment group in a clinical trial. The hazard is the probability that the adverse event will occur in a defined time interval. For example, Nissen et al¹ could measure the number of patients who experience MACE while taking naltrexone-bupropion during week 8 of the study and calculate the risk that an individual patient will experience MACE during week 8, assuming that the patient has not had MACE before week 8. This concept of a discrete hazard rate can be extended to a hazard function, which is generally a continuous curve that describes how the hazard changes over time. The hazard function shows the risk at each point in time and is expressed as a rate or number of events per unit of time.²

Calculating the hazard function using time-to-event observations is challenging because the event of interest is usually not observed in all patients. Thus, the time to the event occurrence for some patients is invisible—or censored—and there is no way to know if the event will occur in the near future, the distant future, or never. Censoring may occur because the patient is lost to follow-up or did not experience the event of interest before the end of the study period. In Nissen et al,¹ only 243 patients experienced MACE before the termination of the study, resulting in 8662 censored observations, meaning there were 8662 patients for whom it is not known when they experienced MACE, if ever. Common nonparametric statistical tests, such as the Wilcoxon rank sum test, could be used to compare the time intervals seen in the 2 groups if the analysis was limited to only the 243 patients who had observed events; however, when censored data are excluded from analysis, the information contained in the experience of the other 8662 patients is lost. While it is unknown when in the future, if ever, these patients will experience an event, the knowledge that these patients did not experience MACE during their participation in the trial is informative. The information contained in censored observations varies: patients whose data are censored early, such as a patient who is lost to follow-up in the first weeks of a study, con-

tribute less information than those who are observed for a long time before censoring. However, all observations provide some information, and to avoid bias, methods of analysis that can accommodate censoring are used for time-to-event studies.

Kaplan-Meier plots and the Cox proportional hazards model are examples of methods for analyzing time-to-event data that account for censored observations. A Kaplan-Meier curve plots the fraction of "surviving" patients (those who have not experienced an event) against time for each treatment group. The height of the Kaplan-Meier curve at the end of each time interval is determined by taking the fraction or proportion of patients who remained event-free at the end of the prior time interval and multiplying that proportion by the fraction of patients who survive the current time interval without experiencing an event. The value of the Kaplan-Meier curve at the end of the current time interval then becomes the starting value for the next time interval. This iterative and cumulative multiplication process begins with the first time interval and continues in a stepwise manner along the Kaplan-Meier curve; the Kaplan-Meier curve is thus sometimes called the "product limit estimate" of the survival curve. Censoring is properly taken into account because only patients still being followed up at the beginning of each time interval are considered in determining the fraction "surviving" at the end of that time interval.³ Figures 2A and 2B in Nissen et al¹ plot the cumulative incidence of MACE in each group vs time, an "upside-down" version of Kaplan-Meier, which provides similar information.

While a Kaplan-Meier plot elegantly represents differences between different groups' survival curves over time, it gives little indication of their statistical significance. The statistical significance of observed differences can be tested with a log-rank test.³ This test, however, does not account for confounding variables, such as differences in patient demographics between groups.

The Cox proportional hazards model both addresses the problem of censoring and allows adjustment for multiple prognostic independent variables, or confounders such as age and sex. The model assumes a "baseline" hazard function exists for individuals whose independent predictor variables are all equal to their reference value. The baseline hazard function is not explicitly defined but is allowed to take any shape. The output of a Cox proportional hazards model is a hazard ratio for each independent predictor variable, which defines how much the hazard is multiplied for each unit change in the variable of interest as compared with the baseline hazard function. Hazard ratios can be calculated for all independent variables, both confounders and intervention variables.

What Are the Limitations of the Proportional Hazards Model?

The Cox proportional hazards model relies on 2 important assumptions. The first is that data censoring is independent of outcome of interest. If the placebo patients in the trial by Nissen et al¹ were both less likely to experience MACE and less likely to follow up with trial investigators because they did not experience weight loss, the probability of censoring and the risk of MACE would be correlated,

threatening the validity of the analysis. The second assumption is that the hazard functions, representing the risk of an event over time, are proportional to each other for all patient groups. In other words, the hazard functions all have the same shape and differ only in overall magnitude; the effect of each independent predictor or confounder is on the overall magnitude of the hazard function. In this trial, it is reasonable to assume that the baseline hazard function for MACE in patients taking placebo looks like a line with a positive slope: age likely increases the hazard of a cardiovascular event. The assumption of proportional hazards means that the hazard function of MACE in patients taking naltrexone-bupropion is assumed to be the baseline hazard multiplied by an unknown, constant value. This assumption would be violated if, for example, patients taking the drug experience an early increase in risk of MACE after initiating treatment as a result of adverse effects but then experience decreased risk over the long-term as they lose weight. In that scenario, the treatment group hazard function would be shaped like a peak with a long tail and would not be proportional to the baseline hazard function.

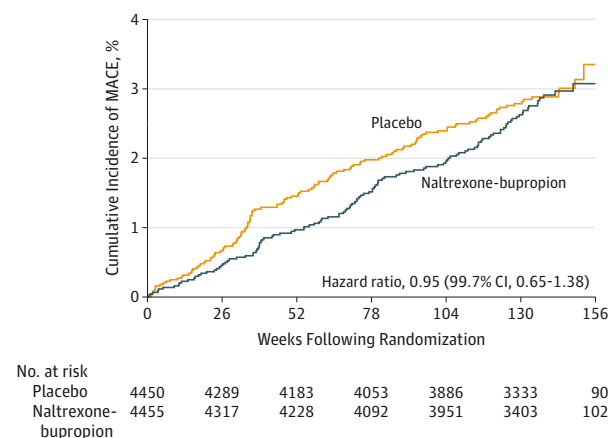
How Should Time-to-Event Findings Be Interpreted in This Particular Study?

The trial was designed as a noninferiority study and statistically powered to assess the null hypothesis that the hazard ratio of naltrexone-bupropion to placebo for MACE is greater than 2.0 at the 25% interim analysis point. Using a Cox proportional hazard model with randomized treatment as a predictor, the estimated hazard ratio was 0.59 (95% CI, 0.39-0.90). It can therefore be concluded that the hazard ratio of MACE associated with the active treatment was less than 2.0. Although it might be tempting to conclude that the hazard ratio is smaller (eg, less than 1.0), the hypothesis testing structure of the noninferiority trial only allowed a rigorous conclusion to be drawn about the hypothesis that the hazard ratio was less than 2.0.

Caveats to Consider When Looking at Results From a Time-to-Event Analysis

Nissen et al¹ used a Cox proportional hazards model to estimate the hazard ratio associated with naltrexone-bupropion compared with placebo for MACE in overweight or obese patients with cardiovas-

Figure. Time to MACE in the Final End-of-Study Analysis



The survival curves cross in this figure from Nissen et al,¹ suggesting that the proportionality assumption may have been violated. MACE indicates major adverse cardiovascular events.

cular risk factors. This trial likely meets the assumptions of the Cox proportional hazards model: the censoring is likely to be independent of hazard, and the hazard functions for all groups are likely to be roughly proportional. Readers should interpret with caution any time-to-event analysis in which the probability of being lost to follow-up or the duration of observation is likely to be correlated with the risk of experiencing an event. Readers should also be cautious in accepting Cox proportional hazards models in which the hazard function of a treatment group is unlikely to be proportional to the baseline hazard. If 2 survival curves cross at any point, such as seen in the far right of Figure 2B in the article by Nissen et al,¹ this might suggest that the hazard ratio between the 2 groups has reversed and the proportionality assumption has been violated (Figure). There are also several diagnostic tests that researchers can use to verify the proportionality assumption, including using Kaplan-Meier curves, testing the significance of time-dependent covariates, and plotting Schoenfeld residuals.⁴ Selection of an appropriate verification method depends on the types of covariates used in the Cox proportional hazards model.

ARTICLE INFORMATION

Author Affiliations: Department of Emergency Medicine, Harbor-UCLA Medical Center, Torrance, California (Tolles, Lewis); Los Angeles Biomedical Research Institute, Torrance, California (Tolles, Lewis); David Geffen School of Medicine at UCLA, Los Angeles, California (Tolles, Lewis); Berry Consultants LLC, Austin, Texas (Lewis).

Corresponding Author: Roger J. Lewis, MD, PhD, Department of Emergency Medicine, Harbor-UCLA Medical Center, Bldg D9, 1000 W Carson St, Torrance, CA 90509 (roger@emedharbor.edu).

Section Editors: Roger J. Lewis, MD, PhD,

Department of Emergency Medicine, Harbor-UCLA Medical Center and David Geffen School of Medicine at UCLA; and Edward H. Livingston, MD, Deputy Editor, *JAMA*.

Conflict of Interest Disclosures: Both authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

REFERENCES

1. Nissen SE, Wolski KE, Prcela L, et al. Effect of naltrexone-bupropion on major adverse cardiovascular events in overweight and obese

patients with cardiovascular risk factors: a randomized clinical trial. *JAMA*. doi:10.1001/jama.2016.1558.

2. Lee ET. *Statistical Methods for Survival Analysis*. 2nd ed. New York, NY: John Wiley & Sons; 1992.

3. Young KD, Menegazzi JJ, Lewis RJ. Statistical methodology: IX, Survival analysis. *Acad Emerg Med*. 1999;6(3):244-249.

4. Hess KR. Graphical methods for assessing violations of the proportional hazards assumption in Cox regression. *Stat Med*. 1995;14(15):1707-1723.

The Stepped-Wedge Clinical Trial Evaluation by Rolling Deployment

Susan S. Ellenberg, PhD

Cluster randomized trials are studies in which groups of individuals, for example those associated with specific clinics, families, or geographical areas, are randomized between an experimental intervention and a control.¹ A stepped-



Related article [page 567](#)

wedge design is a type of cluster design in which the clusters are randomized to the order in which they receive the experimental regimen. All clusters begin the study with the control intervention, and by the end of the trial (assuming no unexpected and unacceptable safety issues arise), all clusters are receiving the experimental regimen.

Use of the Method

Why Is a Stepped-Wedge Clinical Trial Design Used?

Cluster randomized trials have been performed for many decades, even centuries,² but the statistical underpinnings of such designs have been worked out only relatively recently.^{3,4} The primary motivation for a cluster design is to study treatments that can be delivered only in a group setting (eg, an educational approach in a classroom setting) or to avoid contamination in the delivery of each regimen (eg, a behavioral intervention that could be delivered individually but in settings in which those randomized to different approaches are in close contact with each other and might learn about and then adopt the alternative regimen).¹ Clusters are typically identified prospectively and randomized to receive the experimental or control intervention. However, there are exceptions, such as the ring vaccination trial conducted during the 2014-2015 Ebola epidemic, in which clusters were defined around newly identified cases.⁵

If a cluster randomized trial is deemed necessary or desirable in a specific setting, but resource limitations permit only a gradual implementation of the experimental regimen, a stepped-wedge design may be considered as the fairest way to determine which clusters receive the experimental regimen earlier and which later. Stepped-wedge designs have benefits similar to those of crossover trials because outcomes within a cluster may be compared between the time intervals in which a cluster received the control and the experimental interventions. This controls for the unique characteristics of the cluster when making the treatment comparison. One attractive aspect of stepped-wedge designs is that all participants in all clusters ultimately receive the experimental regimen, thereby ensuring that all participants have an opportunity to potentially benefit from the intervention. This can be advantageous when strong beliefs exist regarding the efficacy of a treatment regimen. When limited resources preclude making the treatment regimen widely available from the start, the use of randomization to determine which clusters get early access to the treatment regimen may appeal to participants' sense of fairness.

Description of the Stepped-Wedge Clinical Trial Design

Important considerations in designing a stepped-wedge trial include the number of clusters, the number of "steps" (time points at which the changeovers from control to intervention occur), the duration of treatment at each step, and the balance of prognostic characteristics across the clusters receiving the intervention at each step. The required sample size (total number of participants) to achieve a given level of power decreases as the numbers of clusters and steps increase. Maximum power for a given number of clusters is achieved when each cluster has its own step, but more typically multiple clusters are randomized to change at the same time to limit trial duration.⁶ The risk of bias decreases as the number of clusters increases, as more clusters improve the likelihood of achieving similar prognoses across clusters, and as the trial duration decreases, reducing the effect of temporal trends.

Limitations of the Stepped-Wedge Design

As with cluster randomized designs generally, stepped-wedge designs require larger sample sizes, often much larger, than would be required for randomized trials in which individual study participants are randomized to receive the experimental or control intervention. Efficiency is reduced because of the need to account for the similarities among participants within a given cluster; ie, the extent to which individuals within a cluster are more alike than they are similar to the study population as a whole. Consequently, each individual in a cluster provides less information about the study findings than would occur if the randomization had been by individual. For example, suppose the outcome of a trial was 1-year survival, and in 1 cluster the prognosis of participants was so good that every participant in the cluster was certain to survive at least 1 year. Then the information from that cluster is the same whether there are 100 participants or only 1 participant. When participants are randomized individually, the factors that influence outcomes are balanced within each participating site, and in an analysis appropriately stratified by site, the comparisons will not be affected by site differences in prognosis. Even though randomization of clusters is intended to balance prognosis, such balance cannot be ensured with a small number of clusters (eg, 10-20), which is common in many cluster randomized trials. The randomization can be stratified according to characteristics that are considered to relate to prognosis (eg, mean socioeconomic status of cluster participants), but this is often difficult to do precisely. Unless the number of clusters is quite large, stratification by more than 1 or 2 variables is not feasible.

Another limitation of the stepped-wedge design is the potential for confounding by temporal trends. When changes in clinical care are occurring over a short time, comparisons of outcomes between earlier and later periods may be influenced by background changes that affect the outcome of interest irrespective of the

intervention being tested. Another time-dependent phenomenon that can influence stepped-wedge trials is the effect of accumulating experience with the intervention. If more experience enhances the likelihood that the intervention will be successful, participants in clusters randomized earlier in the trial will more likely benefit. Time dependency concerns must be balanced against the advantage that the before-after comparison within clusters balances the unknown as well as the known characteristics of cluster participants. To address the time dependency, the time factor must be accounted for in the analysis.

How Was the Stepped-Wedge Design Used?

In this issue of *JAMA*, Huffman and colleagues⁷ report results of the QUIK trial, an investigation of a quality improvement intervention intended to reduce complications following myocardial infarction. A stepped-wedge design was used rather than a standard cluster randomized design because this approach allowed all the participating hospitals to receive the experimental intervention during the course of the study and also had the advantage of controlling for potential differences in study participant characteristics by comparing outcomes within a cluster during different periods.⁸ The authors did not pursue an individually randomized design, which also would have controlled for both cluster characteristics and temporal trends. Individual randomization for quality improvement interventions would probably not be feasible within individual participating hospitals because the intervention would be difficult to isolate to individual patients. Sixty-three hospitals were included in the study and were randomized in groups of 12 or 13 that would initiate the

intervention at 1 of 4 randomization points. The duration of each of the 4 steps was 4 months. After adjusting for within-hospital clustering and temporal trends, the prognostic characteristics of the trial participants in the 2 treatment groups were similar.

How Should a Stepped-Wedge Clinical Trial Be Interpreted?

Huffman et al did not find a significant benefit of the quality improvement intervention. Although unadjusted analyses did suggest benefit, appropriate statistical analysis adjusting for time trends markedly attenuated the benefit. In this case, it is possible that the quality of care was improving while the study was progressing independent of the study intervention, highlighting the importance of accounting for time trends (clearly shown in Figures 2A and 2B in the article⁷) when analyzing the results of stepped-wedge trials.

Concerns have been raised about the difficulties in obtaining informed consent from patients in stepped-wedge trials.⁹ Obtaining individual informed consent is often difficult in cluster randomized trials because individuals receiving treatment in a particular cluster may not be able to avoid exposure to the intervention assigned to that cluster. In the QUIK trial, consent was not obtained from patients who received the assigned intervention but it was obtained for 30-day follow-up. The investigators noted that this requirement may have introduced selection bias because of refusals by some participants.

Stepped-wedge clinical trials offer a way to evaluate an intervention in a system in which the ultimate goal is to implement the intervention at all sites yet retain the ability to objectively evaluate the intervention's efficacy.

ARTICLE INFORMATION

Author Affiliation: Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia.

Corresponding Author: Susan S. Ellenberg, PhD, Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, 423 Guardian Dr, Blockley 611, Philadelphia, PA 19104 (sellenbe@penmedicine.upenn.edu).

Section Editors: Roger J. Lewis, MD, PhD, Department of Emergency Medicine, Harbor-UCLA Medical Center and David Geffen School of Medicine at UCLA; and Edward H. Livingston, MD, Deputy Editor, *JAMA*.

Conflict of Interest Disclosures: The author has completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

REFERENCES

1. Meurer WJ, Lewis RJ. Cluster randomized trials: evaluating treatments applied to groups. *JAMA*. 2015;313(20):2068-2069.
2. Moberg J, Kramer M. A brief history of the cluster randomised trial design. *J R Soc Med*. 2015;108(5):192-198.
3. Cornfield J. Randomization by group: a formal analysis. *Am J Epidemiol*. 1978;108(2):100-102.
4. Donner A, Birkett N, Buck C. Randomization by cluster: sample size requirements and analysis. *Am J Epidemiol*. 1981;114(6):906-914.
5. Henao-Restrepo AM, Camacho A, Longini IM, et al. Efficacy and effectiveness of an rVSV-vectored vaccine in preventing Ebola virus disease: final results from the Guinea ring vaccination, open-label, cluster-randomised trial (Ebola Ça Suffit!). *Lancet*. 2017;389(10068):505-518.
6. Baio G, Copas A, Ambler G, Hargreaves J, Beard E, Omar RZ. Sample size calculation for a stepped wedge trial. *Trials*. 2015;16:354.
7. Huffman MD, Mohanan PP, Devarajan R, et al. Effect of a quality improvement intervention on clinical outcomes in patients in India with acute myocardial infarction: the ACS QUIK randomized clinical trial [published February 13, 2018]. *JAMA*. doi:10.1001/jama.2017.21906
8. Huffman MD, Mohanan PP, Devarajan R, et al. Acute coronary syndrome quality improvement in Kerala (ACS QUIK): rationale and design for a cluster-randomized stepped-wedge trial. *Am Heart J*. 2017;185:154-160.
9. Taljaard M, Hemming K, Shah L, Giraudeau B, Grimshaw JM, Weijer C. Inadequacy of ethical conduct and reporting of stepped wedge cluster randomized trials: results from a systematic review. *Clin Trials*. 2017;14(4):333-341.

Mendelian Randomization

Connor A. Emdin, DPhil; Amit V. Khera, MD; Sekar Kathiresan, MD

Mendelian randomization uses genetic variants to determine whether an observational association between a risk factor and an outcome is consistent with a causal effect.¹ Mendelian randomization relies on the natural, random assortment of genetic variants during meiosis yielding a random distribution of genetic variants in a population.¹ Individuals are naturally assigned at birth to inherit a genetic variant that affects a risk factor (eg, a gene variant that raises low-density lipoprotein [LDL] cholesterol levels) or not inherit such a variant. Individuals who carry the variant and those who do not are then followed up for the development of an outcome of interest. Because these genetic variants are typically unassociated with confounders, differences in the outcome between those who carry the variant and those who do not can be attributed to the difference in the risk factor. For example, a genetic variant associated with higher LDL cholesterol levels that also is associated with a higher risk of coronary heart disease would provide supportive evidence for a causal effect of LDL cholesterol on coronary heart disease.



[Author Audio Interview](#)



[CME Quiz](#)

One way to explain the principles of mendelian randomization is through an example: the study of the relationship of high-density lipoprotein (HDL) cholesterol and triglycerides with coronary heart disease. Increased HDL cholesterol levels are associated with a lower risk of coronary heart disease, an association that remains significant even after multivariable adjustment.² By contrast, an association between increased triglyceride levels and coronary risk is no longer significant following multivariable analyses. These observations have been interpreted as HDL cholesterol being a causal driver of coronary heart disease, whereas triglyceride level is a correlated bystander.² To better understand these relationships, researchers have used mendelian randomization to test whether the observational associations between HDL cholesterol or triglyceride levels and coronary heart disease risk are consistent with causal relationships.³⁻⁵

Use of the Method

Why Is Mendelian Randomization Used?

Basic principles of mendelian randomization can be understood through comparison with a randomized clinical trial. To answer the question of whether raising HDL cholesterol levels with a treatment will reduce the risk of coronary heart disease, individuals might be randomized to receive a treatment that raises HDL cholesterol levels and a placebo that does not have this effect. If there is a causal effect of HDL cholesterol on coronary heart disease, a drug that raises HDL cholesterol levels should eventually reduce the risk of coronary heart disease. However, randomized trials are costly, take a great deal of time, and may be impractical to carry out, or there may not be an intervention to test a certain hypothesis, limiting the number of clinical questions that can be answered by randomized trials.

What Are the Limitations of Mendelian Randomization?

Mendelian randomization rests on 3 assumptions: (1) the genetic variant is associated with the risk factor; (2) the genetic variant is not associated with confounders; and (3) the genetic variant influences the outcome only through the risk factor. The second and third assumptions are collectively known as independence from pleiotropy. *Pleiotropy* refers to a genetic variant influencing the outcome through pathways independent of the risk factor. The first assumption can be evaluated directly by examining the strength of association of the genetic variant with the risk factor. The second and third assumptions, however, cannot be empirically proven and require both judgment by the investigators and the performance of various sensitivity analyses.

If genetic variants are pleiotropic, mendelian randomization studies may be biased. For example, if genetic variants that increase HDL cholesterol levels also affect the risk of coronary heart disease through an independent pathway (eg, by decreasing inflammation), a causal effect of HDL cholesterol on coronary heart disease may be claimed when the true causal effect is due to the alternate pathway.

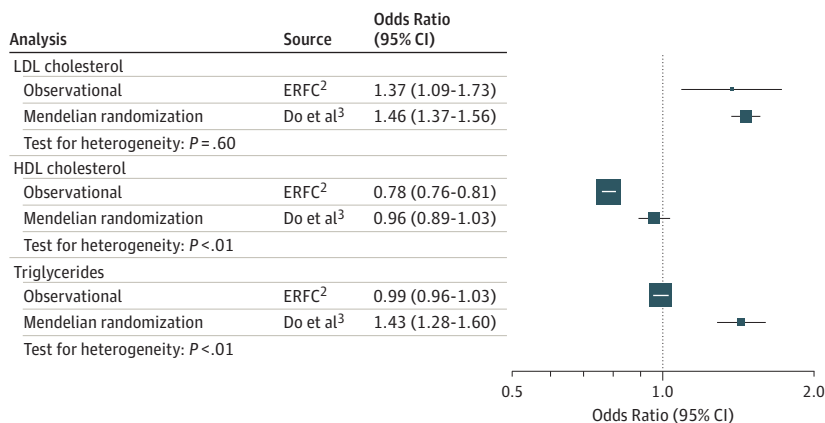
Another limitation is statistical power. Determinants of statistical power in a mendelian randomization study include the frequency of the genetic variant(s) used, the effect size of the variant on the risk factor, and study sample size. Because any given genetic variant typically explains only a small proportion of the variance in the risk factor, multiple variants are often combined into a polygenic risk score to increase statistical power.

How Did the Authors Use Mendelian Randomization?

In a previous report in *JAMA*, Frikke-Schmidt et al⁴ initially applied mendelian randomization to HDL cholesterol and coronary heart disease using gene variants in the *ABCA1* gene. When compared with noncarriers, carriers of loss-of-function variants in the *ABCA1* gene displayed a 17-mg/dL lower HDL cholesterol level but did not have an increased risk of coronary heart disease (odds ratio, 0.93; 95% CI, 0.53-1.62). The observed 17-mg/dL decrease in HDL cholesterol level is expected to increase coronary heart disease by 70% and this study had more than 80% power to detect such a difference; thus, the lack of a genetic association of *ABCA1* gene variants and coronary heart disease was unlikely to be due to low statistical power. These data were among the first to cast doubt on the causal role of HDL cholesterol for coronary heart disease. In other mendelian randomization studies, genetic variants that raised HDL cholesterol levels were not associated with reduced risk of coronary heart disease, a result consistent with HDL cholesterol as a noncausal factor.⁵

Low HDL cholesterol levels track with high plasma triglyceride levels, and triglyceride levels reflect the concentration of triglyceride-rich lipoproteins in blood. Using multivariable mendelian randomization, Do et al³ examined the relationship among correlated risk factors such as HDL cholesterol and triglyceride levels. In an

Figure. Comparison of Observational Estimates and Mendelian Randomization Estimates of the Association of Low-Density Lipoprotein (LDL) Cholesterol, High-Density Lipoprotein (HDL) Cholesterol, and Triglycerides With Coronary Heart Disease



Observational estimates are derived from the Emerging Risk Factors Collaboration (ERFC).² Mendelian randomization estimates are derived from Do et al³ based on an analysis of 185 genetic variants that alter plasma lipids and mutually adjusted for other lipid fractions (eg HDL cholesterol and triglycerides for LDL cholesterol). A formal test of heterogeneity (Cochran Q test) shows that the observational and mendelian randomization causal estimates are consistent for LDL cholesterol but not so for HDL cholesterol or triglycerides.

analysis of 185 polymorphisms that altered plasma lipids, a 1-SD increase in HDL cholesterol level (approximately 14 mg/dL) due to genetic variants was not associated with risk of coronary heart disease (odds ratio, 0.96; 95% CI, 0.89-1.03; Figure). In contrast, a 1-SD increase in triglyceride level (approximately 89 mg/dL) was associated with an elevated risk of coronary heart disease (odds ratio, 1.43; 95% CI, 1.28-1.60). LDL cholesterol and triglyceride-rich lipoprotein levels, but not HDL cholesterol level, may be the causal drivers of coronary heart disease risk as demonstrated by these mendelian randomization studies.

Caveats to Consider When Evaluating Mendelian Randomization Studies

The primary concern when evaluating mendelian randomization studies is whether genetic variants used in the study are likely to be pleiotropic. Variants in a single gene that affects an individual risk factor are most likely to affect the outcome only through the risk factor and not have pleiotropic effects. For example, variants in *CRP*,

the gene encoding C-reactive protein, have been used in a mendelian randomization study to exclude a direct causal effect of C-reactive protein on coronary heart disease.⁶ However, variants in single genes that encode a risk factor of interest are often not available. In these cases, pleiotropy can be examined by testing whether the gene variants used are associated with known confounders such as diet, smoking, and lifestyle factors.⁷ More advanced statistical techniques, including median regression⁸ and use of population-specific instruments,⁷ have recently been proposed to protect against pleiotropic variants biasing results.

A second concern relates to whether the mendelian randomization study has adequate statistical power to detect an association. Consequently, an estimate from a mendelian randomization study that is nonsignificant should be accompanied by a power analysis based on the strength of the genetic instrument and the size of the study. Furthermore, mendelian randomization estimates should be compared with results from traditional observational analyses using a formal test for heterogeneity.

ARTICLE INFORMATION

Author Affiliations: Center for Genomic Medicine, Massachusetts General Hospital, Harvard Medical School, Boston (Emdin, Khera, Kathiresan); Cardiovascular Disease Initiative, Broad Institute, Cambridge, Massachusetts (Emdin, Khera, Kathiresan).

Corresponding Author: Sekar Kathiresan, MD, Center for Genomic Medicine, Massachusetts General Hospital, 185 Cambridge St, CPZN 5.830, Boston, MA 02114 (skathiresan1@mgh.harvard.edu).

Section Editors: Roger J. Lewis, MD, PhD, Department of Emergency Medicine, Harbor-UCLA Medical Center and David Geffen School of Medicine at UCLA; and Edward H. Livingston, MD, Deputy Editor, *JAMA*.

Conflict of Interest Disclosures: All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

REFERENCES

- Smith GD, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol*. 2003;32(1):1-22.
- Di Angelantonio E, Sarwar N, Perry P, et al; Emerging Risk Factors Collaboration. Major lipids, apolipoproteins, and risk of vascular disease. *JAMA*. 2009;302(18):1993-2000.
- Do R, Willer CJ, Schmidt EM, et al. Common variants associated with plasma triglycerides and risk for coronary artery disease. *Nat Genet*. 2013;45(11):1345-1352.
- Frikke-Schmidt R, Nordestgaard BG, Stene MCA, et al. Association of loss-of-function mutations in the *ABCA1* gene with high-density lipoprotein cholesterol levels and risk of ischemic heart disease. *JAMA*. 2008;299(21):2524-2532.
- Voight BF, Peloso GM, Orho-Melander M, et al. Plasma HDL cholesterol and risk of myocardial

infarction: a mendelian randomisation study. *Lancet*. 2012;380(9841):572-580.

6. Zacho J, Tybjaerg-Hansen A, Jensen JS, Grande P, Silleesen H, Nordestgaard BG. Genetically elevated C-reactive protein and ischemic vascular disease. *N Engl J Med*. 2008;359(18):1897-1908.

7. Emdin CA, Khera AV, Natarajan P, et al. Genetic association of waist-to-hip ratio with cardiometabolic traits, type 2 diabetes, and coronary heart disease. *JAMA*. 2017;317(6):626-634.

8. Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genet Epidemiol*. 2016;40(4):304-314.

JAMA Guide to Statistics and Methods

Bayesian Analysis: Using Prior Information to Interpret the Results of Clinical Trials

Melanie Quintana, PhD; Kert Viele, PhD; Roger J. Lewis, MD, PhD

In this issue of JAMA, Laptook et al¹ report the results of a clinical trial investigating the effect of hypothermia administered between 6 and 24 hours after birth on death and disability from hypoxic-ischemic encephalopathy (HIE).



Related article [page 1550](#)

Hypothermia is beneficial for HIE when initiated within 6 hours of birth but administering hypothermia that soon after birth is impractical.² The study by Laptook et al¹ addressed the utility of inducing hypothermia 6 or more hours after birth because this is a more realistic time window given the logistics of providing this therapy. Performing this study was difficult because of the limited number of infants expected to be enrolled. To overcome this limitation, the investigators used a Bayesian analysis of the treatment effect to ensure that a clinically useful result would be obtained even if traditional approaches for defining statistical significance were impractical. The Bayesian approach allows for the integration or updating of prior information with newly obtained data to yield a final quantitative summary of the information. Laptook et al¹ considered several options for the representation of prior information—termed neutral, skeptical, and optimistic priors—generating different final summaries of the evidence.

Prior Information

What Is Prior Information?

Prior information is the evidence or beliefs about something that exist prior to or independently of the data to be analyzed. The mathematical representation of prior information (eg, of beliefs regarding the likely efficacy of hypothermia for HIE 6-24 hours after birth) must summarize both the known information and the remaining uncertainty. Some prior information is quite strong, such as data from many similar patients, and might have little remaining uncertainty or it can be weak or uninformative with substantial uncertainty.

Clinicians routinely interpret the results of a new study in the context of prior work. Are the new results consistent? How can new information be synthesized with the old? Often this synthesis is done by clinicians when they consider the totality of evidence used to treat patients or interpret research studies.

Prior information may be formally incorporated in trial analysis using Bayes theorem, which provides a mechanism for synthesizing information from multiple sources.^{3,4} Clear specification of the prior information used and assumptions made need to be reported in the article or appendix to allow transparency in the analysis and reporting of outcomes.

Why Is Prior Information Important?

When large quantities of patient outcome data are available, traditional non-Bayesian (frequentist) and Bayesian approaches for quantifying observed treatment effects will yield similar results because the contribution of the observed data will outweigh that of the prior

information. This is not the case for evaluating HIE treatments because very few neonates are affected. Despite a large research network, Laptook et al¹ were only able to enroll 168 eligible newborns in 8 years.

Prior information facilitates more efficient study design, allowing stronger, more definitive conclusions without requiring additional patients to be included in the study or analysis. As such, the use of prior information is particularly relevant and important for the study of rare diseases where patient resources are limited.

Prior information can take a number of forms. For example, for binary outcomes, the knowledge that an adverse outcome occurs in 15% to 40% of cases is worth the equivalent of having to enroll 30 or more patients into the trial (depending on the certainty attached to this knowledge). Another form of prior information could be beliefs held regarding the effect of a delay beyond 6 hours in instituting therapeutic hypothermia, ie, that the treatment effect at 7 hours is similar to that at 6 hours and the longer it takes to begin treatment, the less effective the treatment is likely to be.

Limitations of Prior Information

Prior information is a form of assumption. As with any assumption, incorrect prior information can result in invalid or misleading conclusions. For instance, if prior information used the assumption that hypothermia becomes less effective with increasing postnatal age and, in fact, waiting until 12 to 24 hours was associated with the greatest benefit, the resulting inferences would likely be incompatible with the data, less accurate, or biased. If the statistical model uses prior information derived from neonates 0 to 6 hours old in evaluating the treatment effect in neonates 6 to 24 hours of age, and is based on the assumption that the patients respond similarly, the results may be biased or less accurate if the 2 age groups actually respond differently to treatment.

These assumptions can be assessed. Just as the modeling assumptions made in logistic regression can be checked through goodness-of-fit tests,⁵ there are tests that can be used to verify agreement between prior and current data. More importantly, some methods for incorporating prior information can explicitly adjust to conflict between the prior and the data, decreasing the reliance on prior information when the new data appear to be inconsistent with the proposed prior information.⁶

How Was Prior Information Used?

Laptook et al¹ incorporated prior information by allowing for the outcome to vary across time windows of 6 to 12 hours and 12 to 24 hours and prespecifying 3 separate prior distributions on the overall treatment effect (Description of Bayesian Analyses and Implementation Details section of the eAppendix in Supplement 2). The neutral prior assumes that the treatment effect diminishes completely after 6 hours, the enthusiastic prior assumes that effect does not

diminish at all after 6 hours, and the skeptical prior assumes that the treatment is detrimental after 6 hours. Primary results are presented based on the neutral prior and, as such, the authors' approach is transparent and easily interpretable. The authors found a 76% probability of benefit with the neutral prior, a 90% probability of benefit with the enthusiastic prior, and a 73% probability of benefit with the skeptical prior.¹

An alternative to this approach might include specifying a model that relates postnatal age at the start of therapeutic hypothermia to the magnitude of the treatment effect, assuming that the effect does not increase over time. This model would explicitly account for a possible decrease in treatment benefit with increasing age at initiation, while still allowing the effect at each age to inform the effects at other ages. Additionally, this model could be heavily informed or anchored in the 0 to 6-hour range using data from previous studies.² With this anchor, inferences would be improved across the range of 6 to 24 hours, with a particular increase in pre-

cision for the time intervals closer to 6 hours. This may have allowed more definitive conclusions to be drawn from the same set of data.

How Should the Trial Results Be Interpreted in Light of the Prior Information?

Laptook et al¹ used a prespecified Bayesian analysis, using prior information, to allow quantitatively rigorous conclusions to be drawn regarding the probability that therapeutic hypothermia is effective 6 to 24 hours after birth in neonates with HIE. Conclusions of the analysis were given as probabilities that benefit exists. For example, the statement that there is "a 76% probability of any reduction in death or disability, and a 64% probability of at least 2% less death or disability" are easily understood by clinicians and can be used to inform clinical care. The use of several options for prior information allows clinicians with different perspectives to have the data interpreted over a range of prior beliefs.

ARTICLE INFORMATION

Author Affiliations: Berry Consultants LLC, Austin, Texas (Quintana, Viele, Lewis); Department of Emergency Medicine, Harbor-UCLA Medical Center, Los Angeles, California (Lewis); Los Angeles Biomedical Research Institute, Torrance, California (Lewis); David Geffen School of Medicine at UCLA, Los Angeles, California (Lewis).

Corresponding Author: Roger J. Lewis, MD, PhD, Department of Emergency Medicine, Harbor-UCLA Medical Center, Bldg D9, 1000 W Carson St, Torrance, CA 90509 (roger@emedharbor.edu).

Section Editors: Roger J. Lewis, MD, PhD, Department of Emergency Medicine, Harbor-UCLA Medical Center and David Geffen School of Medicine at UCLA; and Edward H. Livingston, MD, Deputy Editor, *JAMA*.

Conflict of Interest Disclosures: All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

REFERENCES

1. Laptook AR, Shankaran S, Tyson JE, et al; Eunice Kennedy Shriver National Institute of Child Health and Human Development Neonatal Research Network. Effect of therapeutic hypothermia initiated after 6 hours of age on death or disability among newborns with hypoxic-ischemic encephalopathy: a randomized clinical trial. *JAMA*. doi:10.1001/jama.2017.14972
2. Jacobs SE, Morley CJ, Inder TE, et al; Infant Cooling Evaluation Collaboration. Whole-body hypothermia for term and near-term newborns with hypoxic-ischemic encephalopathy:

a randomized controlled trial. *Arch Pediatr Adolesc Med*. 2011;165(8):692-700.

3. Food and Drug Administration. Guidance for the use of Bayesian statistics in medical device clinical trials. <https://www.fda.gov/MedicalDevices/ucm071072.htm>. Published February 5, 2010. Accessed September 20, 2017.

4. Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Chichester, England: Wiley; 2004.

5. Meurer WJ, Tolles J. Logistic regression diagnostics: understanding how well a model predicts outcomes. *JAMA*. 2017;317(10):1068-1069.

6. Viele K, Berry S, Neuenschwander B, et al. Use of historical control data for assessing treatment effects in clinical trials. *Pharm Stat*. 2014;13(1):41-54.