Clinical Infectious Diseases
CORRESPONDENCE

Correct Calculation of Confidence Interval for Proportion of Superior Comparisons Between Desirability of Outcome Ranking Scores

To the Editor—We would like to raise your readers' attention about a methodological error we have found in a 2015 publication in this journal by Evans et al [1]. This article, which was an invited commentary and is the subject of a laudatory editorial in the same issue [2], has become the key reference for publications using desirability of outcome ranking (DOOR) and response adjusted for duration antibiotic risk (RADAR) methodology. By calculating a confidence interval for a binomial proportion, treating each pairwise comparison as independent, the authors produced very narrow and incorrect confidence intervals for their test statistic.

Multiple articles containing this error and citing Evans et al have now been published, including 2 in first the 4 months of 2022 alone [3, 4]. One of these is published in this journal. Given that Evans et al is the original description using DOOR terminology and is widely cited, we feel that an explanation of the error, as well as the correct way of calculating confidence intervals for DOOR score comparisons, should be published. We outline a method below that is easily implemented. We infer that the authors are aware of their error as subsequent articles by the same author(s) contain correctly calculated confidence intervals [5].

Consider a trial in which n patients are assigned "new" treatment and m patients are assigned "control" treatment. For brevity, let $(n+m)=N$. Every patient's outcome is ranked in descending order from the best outcome to worst outcome using DOOR/RADAR methodology as described by Evans et al. Average ranks are used for ties. Name these ranks (r_1, r_2,…, r_N).

There are $n \times m$ possible pairwise comparisons between DOOR values for patients in the new treatment group and those in the control group. The test statistic proposed by Evans et al is the proportion of these comparisons in which the new treatment is superior. A shortcut to calculation of the proportion of superior DOOR values is to perform the Wilcoxon rank-sum test on the DOOR values grouped by treatment assignment. This will supply Wilcoxon T and/or Mann-Whitney U. T is the sum of the ranks in the "new treatment" group, and U is the number of these comparisons in which the new treatment is superior. If U is not available it can be calculated from T as:

$$U = T - \frac{1}{2}n(n+1).$$

The test statistic is the proportion of comparisons in which new treatment has a superior DOOR:

$$= \frac{U}{nm}.$$

Based on Fisher's principle of randomization, the variance of this proportion under the null hypothesis (from [6]) is:

$$\frac{\text{Variance of the ranks}}{n \times m \times N} = \frac{1}{N-1}\sum_{k=1}^{N}(r_k - \bar{r})^2 / n \times m \times N.$$

The square root of this value is used as an estimate of the standard error to make a normal approximation to the width of the confidence interval. This holds whether or not ties are present.

Application to the illustrative example in Table 2 from Evans et al [1] gives an identical result of 64.8% for the proportion of comparisons in which the new treatment has a superior DOOR but with a correct 95% confidence interval of 43% to 87%. This compares to the incorrect and unrealistically narrow confidence interval of 57% to 71% supplied by Evans et al. It is important to note that the correct confidence interval includes the null value of 50% and the incorrect confidence interval does not, potentially leading to a false claim of superiority.

The reader may have noted that the "number of comparisons with a superior DOOR" is the Mann-Whitney U statistic described in 1947 [7], albeit with different terminology.

Note

Potential conflicts of interest. The authors: No reported conflicts of interest. All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest.

Mark R. Loewenthal,[1,2] Joshua S. Davis[2,3] and Michael Dymock[4]

[1]School of Medicine and Public Health, University of Newcastle, Newcastle, Australia; [2]Department of Immunology and Infectious Diseases, John Hunter Hospital, Newcastle, Australia; [3]Global and Tropical Health Division, Menzies School of Health and Research, Darwin, Australia; and [4]Wesfarmers Centre for Vaccines and Infectious Diseases, Telethon Kids Institute, University of Western Australia, Perth, Australia

References

1. Evans SR, Rubin D, Follmann D, et al. Desirability of outcome ranking (DOOR) and response adjusted for duration of antibiotic risk (RADAR). Clin Infect Dis **2015**; 61:800–6.
2. Molina J, Cisneros JM. A chance to change the paradigm of outcome assessment of antimicrobial stewardship programs. Clin Infect Dis **2015**; 61:807–8.
3. Molina J, Montero-Mateos E, Praena-Segovia J, et al. Seven-versus 14-day course of antibiotics for the treatment of bloodstream infections by Enterobacterales: a randomized, controlled trial. Clin Microbiol Infect **2022**; 28:550–7.
4. Paez-Vega A, Gutierrez-Gutierrez B, Aguera ML, et al. Immunoguided discontinuation of prophylaxis for cytomegalovirus disease in kidney transplant recipients treated with antithymocyte globulin: a randomized clinical trial. Clin Infect Dis **2022**; 74: 757–65.
5. Doernberg SB, Tran TTT, Tong SYC, et al. Good studies evaluate the disease while great studies evaluate the patient: development and application of a desirability of outcome ranking endpoint for *Staphylococcus aureus* bloodstream infection. Clin Infect Dis **2019**; 68:1691–8.
6. StataCorp. Stata 17 base reference manual. College Station, Texas: Stata Press, **2021**.
7. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat **1947**; 18:50–60.

Correspondence: M. R. Loewenthal, Medical Director, Community and Post-Acute Care, Community and Aged Care Services, Hunter New England Local Health District, 3rd Floor, Newcastle Community Health Centre, 670 Hunter St, Newcastle West, NSW 2302, Australia (mark.loewenthal@hnehealth.nsw.gov.au).

Clinical Infectious Diseases® 2023;76(1):175–6

https://doi.org/10.1093/cid/ciac348

Response to Loewenthal et al

To the Editor—We thank the authors for their thoughtful comments [1] regarding our original paper on the desirability of outcome ranking (DOOR) [2]. We regret the absence of a description of appropriate methodology for confidence interval (CI) estimation for the DOOR probability and the incorrectly reported CI. We agree that CI estimation using methodology for a binomial proportion is incorrect. We had provided initial recommendations regarding CI estimation using the bootstrap in a subsequent paper [3] and applied and noted this approach in the first [4] and future studies that we conducted using the DOOR.

We had compared the performance characteristics of various alternative methodologies for CI estimation in subsequent studies. The method proposed by the authors of the letter performs well when there are few ties when making pairwise comparisons, which may occur in settings with tiebreakers, or when analyzing continuous data. When calculating the variance, the method does not account for ties and is derived under the null hypothesis and thus performs less well when there are many ties as may be the case without a tiebreaker, and will generally provide wider CIs than methods than those that derive under the alternative hypothesis. In our studies, the estimator with the best coverage probability properties is that proposed by Halperin et al. [5], now the method that we use for CI estimation. A pseudo score approach can be used to improve coverage probabilities when sample sizes are highly imbalanced between arms. A freely available online tool providing CI estimates of the DOOR probability based on the Halperin et al. methodology and providing comprehensive DOOR analyses is in development and will available soon.

DISCLOSURES

S. R. E. reports the following grants or contracts unrelated to this work: National Institute of Allergy and Infectious Diseases/National Institutes of Health (NIAID/NIH) grant funding support for UM1AI104681; Degruyter (Editor-in-Chief: Statistical Communications in Infectious Diseases); and Aceragen funding support. S. R. E. also reports book royalties from Taylor & Francis; consulting fees from Genentech, AstraZeneca, Takeda, Microbiotix, Johnson & Johnson, Endologix, Chemo Centryx, Becton Dickenson, Atricure, Roivant, Neovasc, Nobel Pharma, Horizon, International Drug Development Institute, and SVB Leerink; payment or honoraria for lectures, presentations, speakers bureaus, manuscript writing or educational events from Analgesic, Anesthetic, and Addiction Clinical Trial Translations, Innovations, Opportunities, and Networks (ACTTION); support for attending meetings and/or travel from Food and Drug Administration, Deming Conference on Applied Statistics, Clinical Trial Transformation Initiative, Council for International Organizations of Medical Sciences, International Chinese Statistical Association Applied Statistics Symposium, and Antimicrobial Resistance and Stewardship Conference; participation on a Data Safety Monitoring Board or Advisory Board for NIH, Breast International Group, University of Pennsylvania, Duke University, Roche, Pfizer, Takeda, Akouos, Apellis, Teva, Vir, DayOneBio, Alexion, Tracon, Rakuten, Abbvie, Nuvelution, Clover, FHI Clinical, Lung Biotech, SAB Biopharm, and Candel; and is a board member of the American Statistical Association, Society for Clinical Trials and Frontier Science Foundation.

Note

Potential conflicts of interest. The author: No reported conflicts of interest. The author has submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest.

Scott Evans

Department of Biostatistics and Bioinformatics, Milken Institute School of Public Health, George Washington University, 6110 Executive Blvd, Suite 750, Rockville, Maryland 20852-3943, USA

References

1. Loewenthal MR, Davis JS, Dymock M. Correct calculation of confidence interval for proportion of superior comparisons between desirability of outcome ranking scores. Clin Infect Dis **2023**; 76: 175–6.
2. Evans SR, Rubin D, Follmann D, et al. Desirability of outcome ranking (DOOR) and response adjusted for duration of antibiotic risk (RADAR). Clin Infect Dis **2015**; 61(5):800–6.
3. Evans SR, Follmann D. Using outcomes to analyze patients rather than patients to analyze outcomes: a step toward pragmatism in benefit:risk evaluation. Stat Biopharm Res **2016**; 8(4):386–93.
4. Van Duin D, Lok JJ, Earley M, et al. Colistin vs. ceftazidime-avibactam in the treatment of infections due to carbapenem-resistant Enterobacteriaceae. Clin Infect Dis **2018**; 66(2):163–71.
5. Halperin M, Hamdy MI, Thall PF. Distribution-free confidence intervals for a parameter of Wilcoxon-Mann-Whitney type for ordered categories and progressive censoring. Biometrics;45: 509–21.

Correspondence: S. Evans (sevans@bsc.gwu.edu).

Clinical Infectious Diseases® 2023;76(1):176

https://doi.org/10.1093/cid/ciac351

When Emulating a Trial, Do as the Trialists Do: Missteps in Estimating Relative Effectiveness of a Severe Acute Respiratory Syndrome Coronavirus 2 Vaccine Booster Dose

To the Editor—We read with interest the recent study by Butt et al [1]. This observational study, conducted in the Department of Veterans Affairs