JAMA Guide to Statistics and Methods

Worst-Rank Score Methods—A Nonparametric Approach to Informatively Missing Data

John M. Lachin, ScD

A previous *JAMA* Guide to Statistics and Methods article[1] briefly reviewed nonparametric statistics. Such statistical approaches represent the data using ranks of values rather than the observed values. This provides valid tests of significance, regardless of the underlying distributions of the values and without the need to posit parametric assumptions—thus the term "nonparametric statistics." In the May 26, 2020, issue of *JAMA*, Baxter and colleagues from the N-TA^{3}CT Research Group[2] used a nonparametric analysis known as the *worst-rank score method* in a manner that also captured information from missing data that could result from worsening of the patient's condition.

Use of the Method

Why Is the Worst-Rank Score Method Used?

Virtually all studies experience some missing data.[3] Missing data are considered *missing completely at random* (MCAR) when the missing data are the result of random processes by which some values are observed and others are missing. If the missing data are indeed MCAR (an untestable hypothesis), then an analysis of the observed data using virtually any statistical method will provide an unbiased test. Such missing data are called *noninformative* because it is assumed that the missing data are the result of a random process, and a missing datum conveys no information about what the missing value might be.

However, it is possible that missing data are *informatively missing*, in which case the missing data result from other outcomes that reflect a change in the patient's status, either improvement or deterioration. For example, in a study of congestive heart failure, missing data resulting from the death of a patient due to worsening heart failure would indicate that this patient had a worse outcome than any patient who survived. In such settings Wittes et al[4] had suggested that a rank analysis could readily capture this information by assigning the *worst ranks* to study participants who died. Lachin[5] then described the statistical properties of a worst-rank analysis and showed that a rank test using worst ranks provided an unbiased statistical test of the difference between groups.

Informatively missing data were an anticipated feature of the N-TA^{3}CT study[2] that was conducted to compare the effect of doxycycline vs placebo on aneurysm growth among patients with small infrarenal abdominal aortic aneurysms.[2] The primary outcome was the maximum transverse diameter (MTD) of the aneurysm relative to the initial baseline value after 2 years of treatment. However, it was possible that some patients might die or experience rupture of the aneurysm and would require endovascular repair. For such patients the MTD measurement at 2 years would be missing and would be *informative* about the status of these patients relative to those who completed the study and had 2-year MTD measurement data available. Thus, the statistical analysis plan for the study prespecified that these informatively missing patients would be included in the primary efficacy analysis using worst ranks. This JAMA Guide to Statistics and Methods describes the nonparametric worst-rank analysis.

Table. Ranks With Tied Worst Ranks

Group	Patient MTD rank score					
A	8	7	4	6	11	11
B	2	3	9	1	5	11

Abbreviation: MTD, maximum transverse diameter.

Description of Worst-Rank Score Analysis

In the previous *JAMA* Guide to Statistics and Methods article,[1] a simple rank test was described for 2 groups, one with 4 observations and the other with 5, in which the 9 original values were replaced by their ranks (1 to 9). However, now suppose this study actually started with 12 patients, 6 in each group, in which 2 patients in group A died and 1 in group B died. In a worst-rank analysis, if all that was known is that these 3 patients died, then these 3 deaths would be assigned the average worst rank of 11 (the mean of 10, 11, and 12), as shown in the **Table**.

However, if other information would allow the 3 deaths to be ranked by a measure of severity, then the analysis could also be conducted using *exact worst ranks*. For example, the deaths might be ordered by their survival time, longer being better, so that the first death (in study time since randomization) would receive the rank of 12; the second, 11; and the third, 10.

Missing data also may occur from nonfatal deterioration of the patient's condition. In such a case the exact worst ranks can be assigned based on mechanisms of informative missing other than mortality. For example, the DREAM study[6] assessed the effect of blockade of the renin-angiotensin system using ramipril vs placebo for the prevention of diabetes in patients with cardiovascular disease or hypertension. A key outcome was the 2-hour poststimulus glucose level in an oral glucose tolerance test. However, the test was not conducted after a study participant developed diabetes, so that the 2-hour glucose levels were informatively missing for such individuals. Thus, study participants who had developed diabetes were assigned exact worst ranks based on the days from randomization to the diagnosis of diabetes. The analysis was then conducted using the Wilcoxon nonparametric rank test.

How Was the Worst-Rank Score Analysis Used?

The above methods were used in the N-TA^{3}CT study analysis. The authors used a rank transformation analysis of covariance in which the rank scores for the 2-year MTD values were compared between groups when also adjusting for the rank score of the baseline MTD and sex.[1] The analysis also used worst ranks for patients who died or who underwent surgical repair. From the CONSORT diagram, 225 patients had an observed measurable MTD at 2 years. An additional 22 who underwent surgical repair were assigned worse ranks of 226 to 247, based on the time to repair (226 the shortest, 247 the longest), and the 7 who died were assigned ranks of 248 to 255, based on the time to death.

There were no differences between groups in the distribution of the measurable normal MTD scores at 2 years and no appreciable differences in the incidence of repair (9 vs 13) or death (4 vs 3).

Limitations of the Worst-Rank Score Analysis

One drawback to rank-based methods, and to a worst-rank score analysis, is that the analysis does not provide an estimate of a parametric "effect size" that is a function of an estimate of the difference in the distribution parameters between the 2 groups, such as a mean difference with 95% confidence limits. However, when using a Wilcoxon test, the Mann-Whitney formulation provides an estimate of a useful parameter-free or distribution-free quantity that describes the difference between groups. This analysis is based on an estimate of the probability that a random study participant from group A will have a higher value than a random study participant from group B, designated as P(A>B). This probability estimate is based on the sum of the ranks in group A and thus is equivalent to the Wilcoxon rank sum test. If there are no tied ranks, the expected probability is P(A>B) = ½. However, to allow for ties, the *Mann-Whitney Difference* is computed as P(A>B) − P(B>A), where P(B>A) is the reverse probability (random B greater than a random A).[1] Under the null hypothesis of no difference in the distributions between groups, the Mann-Whitney difference is zero and the Mann-Whitney test equals the Wilcoxon test.

How Should Worst-Rank Score Analysis Be Interpreted?

A rank analysis of measurements under the MCAR assumption provides a robust and powerful test of a difference between groups in the population distributions, regardless of the shape of the distribution within each group. An analysis with worst ranks for the patients who have informatively missing data then tests whether there is a difference between groups in the distribution of either the measured values, or the distribution of the times at which informative events occur that result in missing data for the measured primary outcome, or both. Lachin and others[5,7-9] have shown that the worst-rank analysis has good power to detect group differences in situations in which the ranks of the observed values differ between groups or the worst ranks differ in the same direction, and neither of the two show a difference in the opposite direction.

ARTICLE INFORMATION

Author Affiliation: Biostatistics Center, Department of Biostatistics and Bioinformatics, Milken Institute School of Public Health, The George Washington University, Rockville, Maryland.

Corresponding Author: John M. Lachin, ScD, Biostatistics Center, Department of Biostatistics and Bioinformatics, Milken Institute School of Public Health, The George Washington University, 6110 Executive Blvd, Rockville, MD 20852 (jml@bsc.gwu.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*.

Published Online: September 28, 2020. doi:10.1001/jama.2020.7709

Conflict of Interest Disclosures: None reported.

REFERENCES

1. Lachin JM. Nonparametric statistical analysis. *JAMA*. 2020;323(20):2080-2081. doi:10.1001/jama.2020.5874

2. Baxter BT, Matsumura J, Curci JA, et al; N-TA^{3}CT Investigators. Effect of doxycycline on aneurysm growth among patients with small infrarenal abdominal aortic aneurysms: a randomized clinical trial. *JAMA*. 2020;323(20):2029-2038. doi:10.1001/jama.2020.5230

3. Newgard CD, Lewis RJ. Missing data: how to best account for what is not known. *JAMA*. 2015;314(9):940-941. doi:10.1001/jama.2015.10516

4. Wittes J, Lakatos E, Probstfield J. Surrogate endpoints in clinical trials: cardiovascular diseases. *Stat Med*. 1989;8(4):415-425. doi:10.1002/sim.4780080405

5. Lachin JM. Worst-rank score analysis with informatively missing observations in clinical trials. *Control Clin Trials*. 1999;20(5):408-422. doi:10.1016/S0197-2456(99)00022-7

6. Bosch J, Yusuf S, Gerstein HC, et al; DREAM Trial Investigators. Effect of ramipril on the incidence of diabetes. *N Engl J Med*. 2006;355(15):1551-1562. doi:10.1056/NEJMoa065061

7. McMahon RP, Harrell FE Jr. Power calculation for clinical trials when the outcome is a composite ranking of survival and a nonfatal outcome. *Control Clin Trials*. 2000;21(4):305-312. doi:10.1016/S0197-2456(00)00052-0

8. Matsouaka RA, Betensky RA. Power and sample size calculations for the Wilcoxon-Mann-Whitney test in the presence of death-censored observations. *Stat Med*. 2015;34(3):406-431. doi:10.1002/sim.6355

9. Colantuoni E, Scharfstein DO, Wang C, et al. Statistical methods to compare functional outcomes in randomized controlled trials with high mortality. *BMJ*. 2018;360:j5748. doi:10.1136/bmj.j5748