Statistical Analysis of Aneurysmal Subarachnoid Hemorrhage Trials

William H Shuman (will.shuman@icahn.mssm.edu)
Icahn School of Medicine at Mount Sinai https://orcid.org/0000-0001-5818-3848

Sean N Neifert
Icahn School of Medicine at Mount Sinai

Michael L Martini
Icahn School of Medicine at Mount Sinai

Emily K Chapman
Icahn School of Medicine at Mount Sinai

Alexander J Schupper
Icahn School of Medicine at Mount Sinai

Eric K Oermann
New York University School of Medicine

J Mocco
Icahn School of Medicine at Mount Sinai

Jeroen Boogaarts
Radboud University Nijmegen: Radboud Universiteit

Michael D Cusimano
St Michael's Hospital

Airton Leonardo de Oliveira Manoel
Hospital Israelita Albert Einstein Biblioteca Lieselotte Adler ZL: Hospital Israelita Albert Einstein

Nima Etminan
Medical Faculty Mannheim, University of Heidelberg

Hitoshi Fukuda
Kurashiki Central Hospital: Kurashiki Chuo Byoin

Daniel Hanggi
University Hospital Dusseldorf: Universitatsklinikum Dusseldorf

Blessing N.R. Jaja
St Michael's Hospital

Peter Kirkpatrick
Cambridge University: University of Cambridge

Peter Le Roux
Bassett Healthcare

Benjamin Lo

Lenox Hill Hospital
Stephen Mayer
Westchester Medical Center
Andrew Molyneux
University of Oxford
Audrey Quinn
James Cook University Hospital
Gabriel J.E. Rinkel
UMC Utrecht Brain Center: UMC Utrecht Brain Center Rudolf Magnus
Tom A Schweizer
St Michael's Hospital
Jose I Suarez
Johns Hopkins University
Michael Todd
University of Minnesota Health
James C Torner
The University of Iowa
Walter M van den Bergh
University of Groningen: Rijksuniversiteit Groningen
Mervyn D.I. Vergouwen
UMC Utrecht Brain Center Rudolf Magnus
Sen Yamagata
Kurashiki Central Hospital: Kurashiki Chuo Byoin
George Wong
Prince of Wales Hospital
Robert Loch Macdonald
University of California San Francisco Fresno Center for Medical Education and Research

Research Article

Keywords: aneurysm, subarachnoid hemorrhage, SAH, SAHIT, trials, statistical analysis

Posted Date: May 10th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-494126/v1

Abstract

Background: Many randomized controlled trials (RCT) have assessed new treatments in subarachnoid hemorrhage (SAH), yet most show no treatment efficacy. One explanation is the statistical analysis of the primary endpoint was not as efficient as possible. We reanalyzed SAH RCTs with various statistical tests to determine whether the statistical method affects RCT primary outcome.

Methods: Individual patient data for the primary outcome (Glasgow outcome scale [GOS]) of two SAH RCTs were analyzed using 15 statistical methods. For tests requiring outcome dichotomization, multiple cut-points in the 5-level GOS were assessed. Next, a synthetic dataset generated using random sampling with replacement from ten SAH RCTs was assessed using the same statistical tests. A Friedman test (two-way non-parametric analysis of variance) determined which tests produced the highest average absolute *Z*-values. The number of times each test reported significance of $p<0.05$ across the different datasets was calculated.

Results: Bootstrapping with replacement produced the best-ranking results, followed by three χ^2-tests: one differentiating excellent (GOS=5) from good (GOS=4), poor (GOS=2-3), or dead (GOS=1) outcomes; one differentiating favorable (GOS=4-5) from poor or dead outcomes; and one differentiating favorable (GOS=4-5) from unfavorable outcomes. Each of these reported statistical significance for both RCTs, as did the following ranked tests, respectively: Wilcoxon median test, Student's t-test, ordinal logistic regression, median test, and a chi-square dichotomizing excellent (GOS≥4) and inferior outcomes. Statistical significance for one or neither RCT was reported by two Cochrane-Armitage tests and two logistic regressions with alternate versions of bucketing, the Kolmogorov-Smirnov test and chi-square test differentiating surviving from dead patients. The synthetic dataset returned similar results, with the same nine most and six least efficient tests.

Conclusions: Bootstrapping produced the most efficient results but is time-and resource-intensive. Chi-square tests grouping outcomes into dichotomous or multiple buckets are also efficient, and their ease of use and popularity make them appropriate candidates for statistical analysis in future SAH RCTs.

Introduction

Many randomized clinical trials (RCT) have been conducted on patients with subarachnoid hemorrhage (SAH).[1] Only nimodipine and endovascular coiling have garnered robust support.[2,3] Other changes in management of these patients probably have contributed to improved outcomes, including early aneurysm repair, increased use of endovascular techniques, neurocritical care and better diagnosis of minor cases of SAH.[4,5] Nevertheless, still only 36–55% of patients regain independence and 35% succumb to their illness.[6–8]

There are many possible reasons for the paucity of RCT demonstrating efficacy in SAH. Some include inadequate sample size or effect size of the treatment, adverse effects of the tested treatment, efficacy of

rescue therapy in the placebo groups, lack of efficacy of the tested interventions, and insensitivity or suboptimal statistical analysis of the primary outcome. While studies on statistical power on SAH RCT have been conducted,[9] the history of examining statistical procedures to optimize RCT is comparatively much stronger in the ischemic stroke literature,[10–13] to the point that a collaboration (the Optimising Analysis of Stroke Trials [OAST] Collaboration) has been formed to address the topic. While no such collaboration exists for SAH, the Subarachnoid Hemorrhage International Trialists (SAHIT) repository possesses a large amount of RCT and clinical registry data in SAH, which provides an opportunity to study questions of statistical optimization. Hence, the objective of this study was to compare statistical methods in post-hoc analyses of SAH RCT, closely following the methodology of a previous ischemic stroke study by the OAST Collaboration,[10] to determine methodologies for the optimization of statistical efficiency.

Materials And Methods

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Trial Selection and Data Acquisition

Individual patient data in the SAHIT repository was obtained from two RCTs that demonstrated efficacy of the experimental interventions: the British Aneurysm Nimodipine Trial (BRANT)[2] and the International Subarachnoid Aneurysm Trial (ISAT).[3] These were chosen because they are the only two SAHIT trials to have positive effects on outcome.

Data regarding trial characteristics, patient demographics, patient severity and outcome on the Glasgow Outcome Scale (GOS) were collected for each trial. The time point at which outcome was recorded was three months for BRANT and two months for ISAT.

Synthetic Dataset

A synthetic dataset designed to have a 20% difference in favorable outcome was created using raw patient outcomes data from ten RCTs in the SAHIT repository.[14] The dataset was composed of 2,000 patients randomly selected with replacement. Half of the patients belonged to the experimental intervention cohort of the respective trial with the other half having received a placebo. The synthetic placebo cohort was generated through sampling without replacement. Then the synthetic intervention cohort was generated such that it had a 20% relative greater proportion of patients with favorable outcomes, defined as a GOS score of 4 or 5. For trials that reported outcomes on the modified Rankin Scale (mRS) or the extended GOS (GOSE), scores were converted to the GOS using conversion schemes consistent with those of Olsen[15] and Michaud.[16] Outcomes were from 3 or 6 months after SAH depending on data availability for each trial.

Statistical Tests

Fifteen statistical tests were conducted on the raw data from each trial and the synthetic dataset and compared to assess the statistical significance of the RCT treatment effect. These tests were the same ones used by a similar study that examined ischemic stroke RCTs.[10] While some tests analyzed the raw trial data, certain analyses, such as the chi-square test, required placing patients into discrete bins or categories based on their outcome scores. Additionally, some tests required dichotomization of outcomes into 2 or more groups (“good”/“bad”); these tests were assessed multiple times with different “breakpoints” for dichotomization.

Comparison of Statistical Tests

The significance of each test’s result was determined by the *Z*-value of the output. The absolute values of each *Z*-value were then ranked from highest to lowest, such that the highest Z-value and statistically most significant result received a rank of 1 and the lowest Z-value and least significant result received a rank of 15 (i.e. the lower the rank the more efficient the outcome). A nonparametric 2-way analysis of variance (ANOVA) (Friedman test) was used to determine which test produced the lowest average rank and to assess differences in the results of studied tests. Tests were re-ordered in terms of overall average rank, and the number of times each test reported significant results across the different datasets was calculated. Analyses were carried out in SAS (version 9.4) and statistical significance was defined by $p < 0.05$.

Results

The average rank of effect of each statistical test showed that bootstrapping with a non-parametric Wilcoxon test had the highest rank followed by three chi-square tests with different bucketing criteria (excellent-good-poor-dead, good-poor-dead and good-poor, Table 1**)**. For chi-square tests, those with more categories returned more statistically significant results with lower mean ranks, while those that dichotomized outcomes into two categories received higher mean ranks. Tests that analyzed ordinal variables performed better when analyzing raw outcome data on the 5-point GOS scale than when analyzing data artificially bucketed into categories. For example, The Wilcoxon test, Student’s t-test, median test and ordinal logistic regression all received higher mean ranks than the logistic regression when it was bucketed into two or more categories.

Table 1
Mean Rank of Statistical Tests

Mean Rank	Statistical Test	# Significant
1	Bootstrap	100%
2	Chi-square EGPD	100%
3	Chi-square GPD	100%
4	Chi-square G	100%
5.5	Wilcoxon Test	100%
6	T-Test	100%
6.5	Logistic Ordinal	100%
8.5	Median Test	100%
9.5	Chi-square E	100%
11	Logistic EGPD	50%
12	Cochrane Armitage EGPD	50%
12.5	Chi-square A	0%
12.5	Kolmogorov-Smirnov	50%
12.5	Logistic GPD	50%
13.5	Cochrane Armitage GPD	50%

Abbreviations: EGPD – Excellent vs. Good vs. Poor vs. Dead, GPD – Good vs. Poor vs. Dead, G – Good vs. Poor, E – Excellent vs. Inferior Outcome, A – Alive vs. Dead

When the results of each test were assessed for their statistical significance, nine of 15 tests returned statistically significant results for both trials, and these were also the nine studies with the most efficient average ranks. Five of the remaining studies returned statistically significant results for only one of the two trials. One test, a chi-square test dichotomizing patients into alive-dead categories did not return statistically significant results for either study and received an average rank of 12.5 for the two trials. The results returned by each statistical test also differed significantly (Friedman test, $p < 0.0001$, Fig. 1).

Results for the synthetic dataset found the four most efficient tests (bootstrap followed by three chi-square tests) were the same as those obtained by analysis of BRANT and ISAT (Table 2). Additionally, the ranks for the remaining tests matched closely with the original mean rank list, with the nine most efficient tests and six least efficient tests consisting of the same collections of tests, with minor variations in order. In terms of significance, 13 statistical tests returned statistically significant values for the synthetic dataset, with the two least efficient returning non-significant values.

Table 2
Ranks of Statistical Tests for Synthetic Dataset

Rank[a]	Statistical Test
1	Bootstrap
2	Chi-square EGPD
3	Chi-square GPD
4	Chi-square G
5	Chi-square E
6	Wilcoxon Test
7	Logistic Ordinal
8	T-Test
9	Median Test
10	Logistic EGPD
11	Cochrane Armitage EGPD
12	Kolmogorov-Smirnov
13	Logistic GPD
14	Chi-square A
15	Cochrane Armitage GPD

Abbreviations: EGPD – Excellent vs. Good vs. Poor vs. Dead, GPD – Good vs. Poor vs. Dead, G – Good vs. Poor, E – Excellent vs. Inferior Outcome, A – Alive vs. Dead

[a] *Trials included in the synthetic dataset were ALISAH (Albumin in Subarachnoid Hemorrhage), BRANT (British Aneurysm Nimodipine Trial), EPO/Statin (Acute Erythropoietin/Acute Systemic Pravastatin) trials, HHU (Heinrich Heine University), IHAST (Intraoperative Hypothermia for Aneurysm Surgery Trial), MASH (Magnesium sulfate in Aneurysmal Subarachnoid Hemorrhage), NEWTON-1/2 (Nimodipine Microparticles to Enhance Recovery While Reducing Toxicity), and TIRILAZAD.*

Discussion

Outcomes following SAH have improved over the past few decades. Data showing that mortality has decreased are more convincing whereas improvements in the function of survivors are less well documented.[4] While recognizing improvements in SAH outcome, it is notable that only two of the more than 75 SAH RCTs demonstrated treatment efficacy.[17] Thus, a fundamental question is how to maximize the chances that an RCT of a truly effective treatment actually demonstrates this.

One way to do this may be to analyze the primary outcome in some way other than the most common, simple dichotomous “good” or “bad” method.[18] Herein we assessed raw data from BRANT, ISAT and a synthetic dataset using 15 statistical tests in order to identify a statistical method that might optimize the analysis of outcome in SAH RCT.[10,14]

We found that the bootstrap analysis was the most efficient of the included tests. Bootstrapping involves high-volume resampling with random sampling and data replacement. This high-volume repetition of comparisons may function to increase the magnitude of treatment effect size. A study of ischemic stroke RCTs also found bootstrapping performed relatively well.[10] There are limitations of bootstrapping. It requires large sample sizes in order to ensure similar data distributions between trial groups and minimize type I error.[19] Bootstrapping also is heavily resource and time intensive due to the large extent of resampling required to control error rates.

Three different chi-square tests were the next most efficient tests. This is opposite to the results from the ischemic stroke RCT study.[10] That study reported that when similar tests were conducted on the data from ischemic stroke clinical trials, tests that grouped outcome scores into 2 or more categories, such as these three chi-square tests, did not perform as well as tests that evaluated ordinal data.[10] Rather, we observed that within the different chi-square tests, those that divided outcome measurements into 3 or 4 categories performed better than those with only 2 categories. These findings suggest that the certain characteristics of SAH trial design, of the way trial data is collected, or of the outcomes of SAH patients may benefit from the division of such outcomes into pre-defined categories.

It is likely that characteristics of the RCT data, such as the distribution of outcomes across the ordinal scale, influence the statistical efficiency of different tests. For example, most patients in the RCTs included in this study had GOS outcomes of 4 or 5, so statistical tests that are able to differentiate between these scores may more easily detect significant effects of treatment. The distribution of outcomes for patients with other neurological diseases may differ from those of SAH patients. Regarding outcome distribution, it is important to ensure that the data distribution meets all the assumptions of a statistical test when designing the analysis for a study.

Not all studies with outcome scores divided into categories performed efficiently. An exception was a chi-square test with categories for “alive” or “deceased” outcomes, which did not find a statistically significant difference between cohorts in either trial. This comparison differs from the other chi-square tests in that it analyzed effect of the interventions on mortality. The mortality rates were low, however, in the RCT and larger sample sizes may be needed.

Improved test performance from bucketing into more categories was not seen for all statistical tests. The Cochrane-Armitage (divided into either 3 or 4 buckets) and logistic regressions divided into 3 or 4 buckets performed poorly, each only returning statistically significant results for one of the trials, and all had higher mean-ranks than ordinal logistic regression. Notably, all of these tests failed to reach significance

for BRANT. This could be due to the small sample size of BRANT, which was about a quarter the size of ISAT.

Conclusion

The efficiency of different statistical tests for analysis of SAH RCT varied significantly. However, which method to use is only one consideration in design of SAH RCT. Chi-square analyses that assess outcomes across multiple categories of ordinal outcomes tended to perform more efficiently in the included trials and may help future trials identify effective SAH treatments. A frequent limitation in SAH RCT is inadequate sample size and these results should not be taken as ways to reduce sample size.

Declarations

Ethics approval and consent:

Need for approval was waived as this study utilizes only de-identified datasets without any identifying personal health information.

Consent for publication:

Not applicable.

Availability of Data and Materials:

The datasets analyzed during the current study are not publicly available due to their proprietary nature but are available from the corresponding author on reasonable request.

Competing Interests:

RLM is a consultant for Grace Biotechnology and Idorsia Pharmaceuticals. EKO owns equity in MedAugur and Whiteboard Coordinator, receives consulting fees from Google, and is employed at Merck. JM reports receiving research support from Stryker, Penumbra, Medtronic, and Microvention and is a consultant for Imperative Care, Cerebrotech, Viseon, Endostream, Rebound Therapeutics, and Vastrax.

Funding:

No funding was received for the work done in this study.

Authorship:

WHS, SNN, MLL, EKC and AJS each contributed to the design and draft of the work, and they have approved the submitted version and have agreed both to be personally accountable for the accuracy or integrity of the work. EKO, JM, JB, MDC, ALOM, NE, HF, DH, BNRJ, PK, PLR, BL, SM, AM, AQ, GJER, TAS, JIS, MT, JCT, WMVDB, MDIV, GW, SY each contributed to the acquisition and interpretation of data analyzed in the work and substantially revised it, and they have approved the submitted version and have agreed both to be personally accountable for the accuracy or integrity of the work. RLM contributed to the concept, design, and drafting of the work and substantively revised it, and he has approved the submitted version and has agreed both to be personally accountable for the accuracy or integrity of the work

Acknowledgements:

MDIV reports funding from the Dutch Heart Foundation (clinical established investigator grant number 2018T076).

References

1. Macdonald RL. Delayed neurological deterioration after subarachnoid haemorrhage. *Nature reviews Neurology.* 2014;10(1):44-58.
2. Pickard JD, Murray GD, Illingworth R, et al. Effect of oral nimodipine on cerebral infarction and outcome after subarachnoid haemorrhage: British aneurysm nimodipine trial. *BMJ.* 1989;298(6674):636-642.
3. Molyneux AJ, Kerr RS, Yu LM, et al. International subarachnoid aneurysm trial (ISAT) of neurosurgical clipping versus endovascular coiling in 2143 patients with ruptured intracranial aneurysms: a randomised comparison of effects on survival, dependency, seizures, rebleeding, subgroups, and aneurysm occlusion. *Lancet (London, England).* 2005;366(9488):809-817.
4. Neifert SN, Chapman EK, Martini ML, et al. Aneurysmal Subarachnoid Hemorrhage: the Last Decade. *Translational stroke research.* 2020.
5. Chan V, Lindsay P, McQuiggan J, Zagorski B, Hill Michael D, O'Kelly C. Declining Admission and Mortality Rates for Subarachnoid Hemorrhage in Canada Between 2004 and 2015. *Stroke.* 2019;50(1):181-184.
6. Hop JW, Rinkel GJ, Algra A, van Gijn J. Case-fatality rates and functional outcome after subarachnoid hemorrhage: a systematic review. *Stroke.* 1997;28(3):660-664.
7. Nieuwkamp DJ, Setz LE, Algra A, Linn FH, de Rooij NK, Rinkel GJ. Changes in case fatality of aneurysmal subarachnoid haemorrhage over time, according to age, sex, and region: a meta-analysis. *The Lancet Neurology.* 2009;8(7):635-642.
8. Lovelock CE, Rinkel GJ, Rothwell PM. Time trends in outcome of subarachnoid hemorrhage: Population-based study and systematic review. *Neurology.* 2010;74(19):1494-1501.

9. Kreiter KT, Mayer SA, Howard G, et al. Sample size estimates for clinical trials of vasospasm in subarachnoid hemorrhage. *Stroke.* 2009;40(7):2362-2367.
10. Bath PM, Gray LJ, Collier T, Pocock S, Carpenter J. Can we improve the statistical analysis of stroke trials? Statistical reanalysis of functional outcomes in stroke trials. *Stroke.* 2007;38(6):1911-1915.
11. Bath P, Hogg C, Tracy M, Pocock S. Calculation of numbers-needed-to-treat in parallel group trials assessing ordinal outcomes: case examples from acute stroke and stroke prevention. *International journal of stroke : official journal of the International Stroke Society.* 2011;6(6):472-479.
12. Gray LJ, Bath PM, Collier T. Should stroke trials adjust functional outcome for baseline prognostic factors? *Stroke.* 2009;40(3):888-894.
13. Calculation of sample size for stroke trials assessing functional outcome: comparison of binary and ordinal approaches. *International journal of stroke : official journal of the International Stroke Society.* 2008;3(2):78-84.
14. Pace A, Mitchell S, Casselden E, et al. A subarachnoid haemorrhage-specific outcome tool. *Brain.* 2018;141(4):1111-1121.
15. Olsen A. *Cognitive Control Function and Moderate-to-Severe Traumatic Brain Injury: Functional and Structural Brain Correlates* 2014.
16. Michaud C, Thomas W, McAllen K. Early Pharmacologic Treatment of Delirium May Reduce Physical Restraint Use: A Retrospective Study. *The Annals of pharmacotherapy.* 2013;48.
17. Jaja BNR, Attalla D, Macdonald RL, et al. The Subarachnoid Hemorrhage International Trialists (SAHIT) Repository: Advancing Clinical Research in Subarachnoid Hemorrhage. *Neurocritical Care.* 2014;21(3):551-559.
18. Macdonald RL, Cusimano MD, Etminan N, et al. Subarachnoid Hemorrhage International Trialists data repository (SAHIT). *World neurosurgery.* 2013;79(3-4):418-422.
19. Saha S, Brannath W, Bornkamp B. Testing multiple dose combinations in clinical trials. *Stat Methods Med Res.* 2020;29(7):1799-1817.

Figures

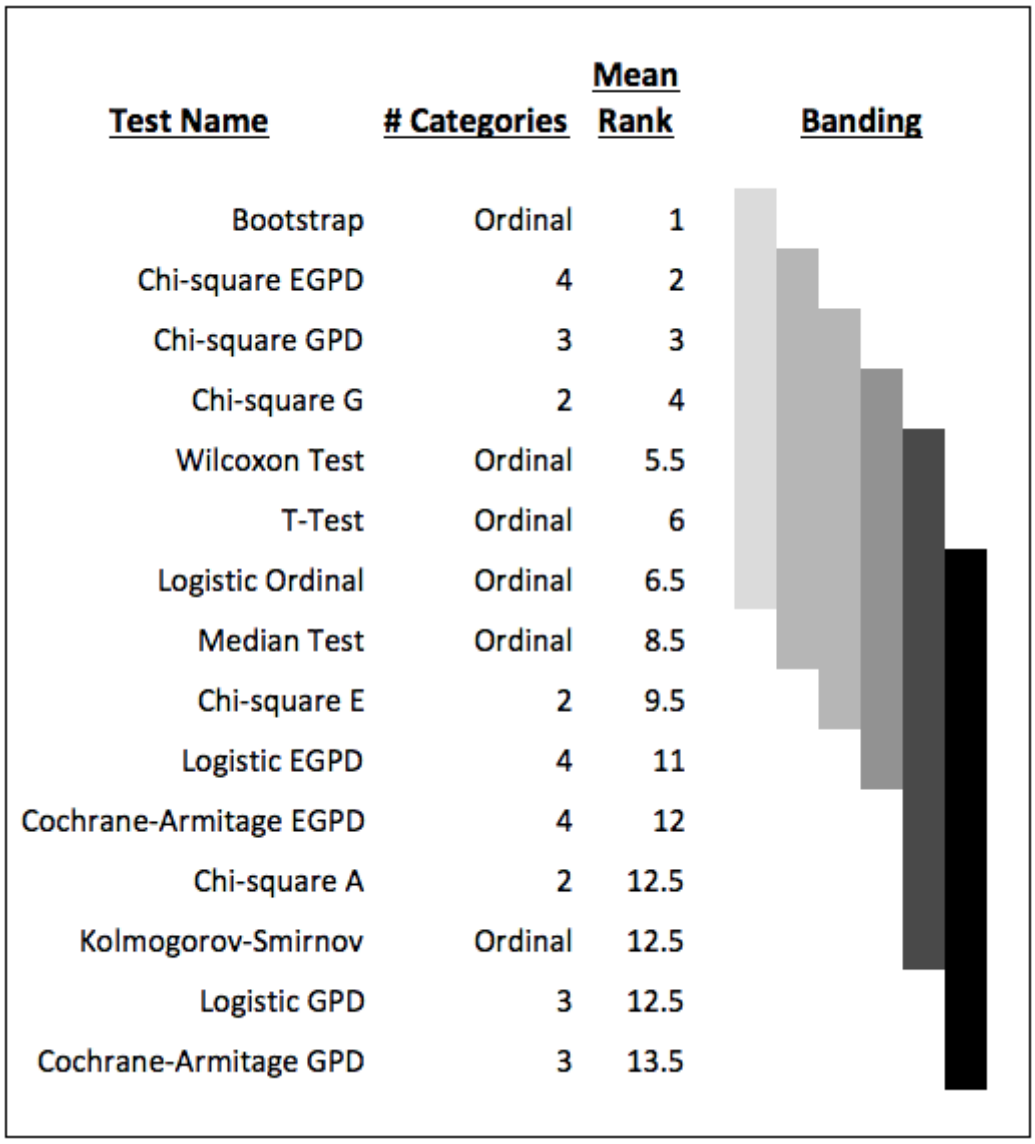

Figure 1

Comparison of Rank Scores for 15 Included Statistical Tests: lower ranks imply the test is more efficient; tests joined by the same band are not statistically significantly different from each other at $P<0.05$