Published in final edited form as:

Mov Disord. 2009 September 15; 24(12): 1732. doi:10.1002/mds.22645.

Using Global Statistical Tests in Long-Term Parkinson's Disease Clinical Trials

Peng Huang, PhD^{1,*}, Christopher G. Goetz, MD², Robert F. Woolson, PhD³, Barbara Tilley, PhD³, Douglas Kerr, MD¹, Yuko Palesch, PhD³, Jordan Elm, MA³, Bernard Ravina, MD, CCRC⁴, Kenneth J. Bergmann, MD⁵, Karl Kieburtz, MD⁴, and The Parkinson Study Group

¹Division of Oncology Biostatistics, Johns Hopkins University, Baltimore, Maryland, USA

²Department of Medicine, Rush University Medical Center, Chicago, Illinois, USA

³Department of Biostatistics, Bioinformatics, and Epidemiology, Medical University of South Carolina, Charleston, South Carolina, USA

⁴Department of Neurology, University of Rochester, Rochester, New York, USA

⁵U.S. Food and Drug Administration, Silver Spring, Maryland, USA

Abstract

Parkinson's disease (PD) impairments are multidimensional, making it difficult to choose a single primary outcome when evaluating treatments to stop or lessen the long-term decline in PD. We review commonly used multivariate statistical methods for assessing a treatment's global impact, and we highlight the novel Global Statistical Test (GST) methodology. We compare the GST to other multivariate approaches using data from two PD trials. In one trial where the treatment showed consistent improvement on all primary and secondary outcomes, the GST was more powerful than other methods in demonstrating significant improvement. In the trial where treatment induced both improvement and deterioration in key outcomes, the GST failed to demonstrate statistical evidence even though other techniques showed significant improvement. Based on the statistical properties of the GST and its relevance to overall treatment benefit, the GST appears particularly well suited for a disease like PD where disability and impairment reflect dysfunction of diverse brain systems and where both disease and treatment side effects impact quality of life. In future long term trials, use of GST for primary statistical analysis would allow the assessment of clinically relevant outcomes rather than the artificial selection of a single primary outcome.

Keywords

multiple outcomes; global treatment effect

Parkinson's disease (PD) is a multifaceted disorder, with motor, cognitive, behavioral, and autonomic features. The standard tool to assess PD has been the Unified Parkinson's Disease Rating Scale (UPDRS),¹ which highly prioritizes motor function evaluation. The scale's revision sponsored by the Movement Disorder Society (MDS),² termed the MDS-UPDRS evaluates non-motor elements of PD more fully, but validation of this scale is still in progress.

Additionally, many PD treatments provoke a spectrum of side effects that can affect tolerability, and therefore overall outcome. These are also not well-rated by currently used instruments.

Traditional statistical analyses are based on a selected primary outcome, usually a change in UDPRS scores related to treatment, with additional secondary outcomes based on other scale results. This strategy implicitly prioritizes the motor signs of PD over the non-motor phenomena and cannot provide a single assessment measure of overall benefit. Furthermore, if the primary outcome is significantly improved, but serious declines occur in other key secondary outcomes, the reporting of the primary outcome effect can be clinically misleading. Likewise, if the primary outcome does not meet statistical significance by itself for improvement, but all the secondary outcomes are positive, there is no widely accepted method to combine these trends to reflect overall improvement in a way that may demonstrate statistical significance.

Various statistical methods have been proposed to analyze clinical trials with multiple outcomes. This paper reviews some of these commonly used multivariate statistical methods, the questions these methods address, and their strengths and limitations in detecting efficacious treatments. In particular, we present the global statistical test (GST), first proposed by O'Brien³ and further extended by others,^{4–17} using data from two PD clinical trials. Whereas GST has been widely used in clinical research on stroke,^{18,19} dermatology,²⁰ multiple sclerosis,²¹ asthma,²² and rheumatoid arthritis,²³ its utility has not been fully explored in PD.

MULTIVARIATE TECHNIQUES

Several strategies have been used to compare treatments when multiple outcomes, each considered important, are targeted. The most common strategy is to choose a single outcome as the primary outcome, leaving all others as secondary outcomes. This strategy, useful when the chosen primary outcome is known to be the most important outcome, may be shortsighted if a treatment effective for multiple outcomes is dismissed simply because the primary outcome measure fails to yield proof of significant efficacy. On the other hand, if the primary outcome shows treatment benefit, this benefit could be difficult to interpret if other important outcomes either show no evidence of benefit or are observed to worsen.

Other frequently employed strategies include the use of a linear combination of several outcomes, multiple tests with adjustment to the overall significance level, omnidirectional tests of any type of treatment difference, and hierarchical models using latent parameters or hyperparameters. These approaches, aimed to test different types of treatment differences, address different scientific questions, i.e., their hypotheses are different. For example, Bonferroni adjustment and its extensions^{24–29} are designed to test whether a treatment has any type of effect—positive or negative. If the medical question is to determine whether one treatment is preferred over the other based on multiple important outcomes, this strategy tends to statistically obscure findings and particularly loses power when the treatment shows consistent improvement across all outcomes.

Less known to many investigators is the global statistical test (GST) which is designed to address whether a treatment is efficacious across different aspects of a condition. The GST efficiently summarizes a treatment's merit when the medical question is complex and even if sample size is relatively small. When a treatment shows improvement on all target outcomes, the GST often has a higher power than tests of single outcomes or other multiple test procedures. As such, GST incorporates the impact of consistent directional change across multiple key target outcomes, even when individual outcomes may not show statistically significant improvement on their own. For example, in a NINDS stroke trial,¹⁸ the t-PA-treated group showed a statistically significant improvement on all four primary outcomes (the Barthel Index,

Modified Rankin Scale, Glasgow Outcome Scale, and National Institutes of Health Stroke Scale), but after Bonferroni adjustment for multiple comparisons, no single outcome reached the required threshold of statistical significance ($P < 0.0125$). In contrast, the P -value from GST with the same data was 0.008, demonstrating a strong evidence of global treatment benefit. On the other hand, GST lowers its power if both strong beneficial and detrimental effects occur among the preselected target outcomes. Such power loss protects against a strong improvement in one area but at the expense of serious decline in another valued target outcome. An overview of these differing methods is summarized in Table 1.

A BRIEF REVIEW OF GST

O'Brien introduced two types of GST: parametric GST and nonparametric GST.³ Several new GSTs have been proposed since O'Brien's seminal work. These tests are most suitable when all outcomes are measured in the same scale.^{30–34} They assume the treatment has a common effect on all outcomes and test whether this common effect is improved. A major barrier to making these GSTs widely adopted in clinical research is the difficulty of interpreting such a common effect because different GSTs have distinct, and generally nonintuitive, interpretations. O'Brien's nonparametric GST does not require a common treatment effect assumption and can be applied to outcomes measured in different scales. Its validity was more broadly established in 2005.¹⁵ A unified interpretation of nonparametric GST and sample size formulae were recently provided through the use of global treatment effect (GTE).¹⁶

The GTE is a value between -1 and 1 with $GTE = 0$ implying no global treatment benefit, $GTE = 1$ implying the treatment is most preferred, and $GTE = -1$ implying the treatment is least preferred. Larger positive GTE values correspond to higher degrees of treatment preference. The GTE, defined through an average of probabilities of treatment benefit on multiple outcomes, plays a similar role as the traditionally used effect size in study design. However, GTE has many advantages over the effect size^{16,17,35,36} and provides a direct method for clinical interpretation. For example, a $GTE = 0.40$ for the total UPDRS implies that a patient will have $(1 + 0.40)/2 = 70\%$ chance of a better UPDRS score on the given treatment. The formula $(1 + GTE)/2$ gives a value between 0 and 100%. The interpretation of GTE is uniform; no matter what measurement scales are used, the GTE is unchanged. This overcomes the weakness of commonly used effect size whose value will be altered if a log-transformation is applied to the data.

TWO EXAMPLES OF GST APPLICATION IN PD

Two Prototypic Data Sets with Different Outcomes

To demonstrate the utility of GST in PD research relative to other techniques, we chose two different PD clinical trials and analyze the multiple outcomes from those studies using different statistical methods. We compare nonparametric GST to other commonly used multivariate techniques: multivariate analysis of variance (MANOVA), Hotellings T^2 test, and Bonferroni adjustment, and specify the technique used in the core analysis published. Data were provided by The Multicenter Effects of Coenzyme Q₁₀ in Early PD study (QE2 trial)³⁷ and The Multicenter Randomized Controlled Trial of Remacemide Hydrochloride as Monotherapy for Parkinson's Disease (RAMP).³⁸ These two trials are chosen because a consistent treatment improvement on multiple outcomes was observed in QE2 trial while both improvement and decline were observed for different outcomes in RAMP trial.

QE2 trial randomized 80 patients to receive either one of the active treatments (doses of 300, 600, and 1200 mg) or placebo with sample sizes of 21, 20, 23, and 16 respectively. The primary analysis using the analysis of covariance has found a linear trend ($P = 0.09$) between the dosage and the mean change in the total UPDRS from baseline to the last visit before starting levodopa

therapy or the month 16 visit. Secondary analyses included analysis of covariance of three subscales of the primary outcome: UPDRS mental, motor, and the activities of daily living (ADL). If all three subscales in UPDRS are considered equally important to defining treatment effect, the use of three multiple primary outcomes with equal weight will more accurately reflect the disease process than the use of total UPDRS which gives higher weights to motor and ADL scores.

RAMP trial randomized 200 patients to receive either one of the active treatments (doses of 150, 300, and 600 mg of remacemide) or placebo with sample sizes of 51, 46, 53, and 50 respectively. The primary efficacy analysis used Bonferroni adjustment to a series of t-tests to compare each treatment group with placebo in the change of the sum of UPDRS motor and ADL subscales from baseline to week 5. Investigators reported no significant treatment difference. Secondary efficacy analysis included the changes in UPDRS subscales for mental, motor, ADL, the Beck Depression Inventory (BDI), and Mini-mental State Exam (MMSE). The investigators concluded that there was no treatment benefit, and that placebo was better than treatment in UPDRS mental subscale. Instead of using the sum of UPDRS motor and ADL as a single primary efficacy measure, we tested treatment efficacy by including additional nonmotor functions (UPDRS mentation, BDI, and MMSE). To illustrate this analysis, we considered all outcomes equally important in the statistical model and did not include weighting of variables.

Data Analysis: GST vs Other Techniques

For both datasets, we used the multiple outcomes considered by the original investigators as primary or secondary and compared different statistical techniques to the nonparametric GST from Huang et al.¹⁵ which considers components in both primary and secondary outcomes. For a comparison, we included univariate Wilcoxon test for each outcome alone, multivariate techniques of Hotellings T^2 test, MANOVA, and Bonferroni adjustment.

Conclusions regarding treatment effect differed depending on which statistical approach was used. In both combined treatment and 1200 mg/day versus placebo comparisons (Table 2), the GST yielded the smallest P -values among all multivariate tests. These results are anchored in the consistent treatment improvements and the similarly positive GTE values across the three outcomes. Bonferroni adjustment yielded the smallest P -value in 300 mg/day versus placebo comparison driven by the single ADL outcome whose GTE value is more than twice larger than the other two outcomes. Among all comparisons, Hotelling's T^2 test was never the most powerful test.

For the RAMP study, where there were combined positive and negative changes, no overall treatment efficacy was found by GST and MANOVA (threshold significance level $\alpha = 0.05$, Table 3). The GST using both RAMP study's primary outcome (Motor + ADL) and secondary outcomes (Mental, BDI, and MMSE) gave P -value = 0.1456 for the combined treatment versus placebo comparison. However, Bonferroni adjustment showed a strong benefit for placebo versus the combined treatment group ($P = 0.0164$), driven by the negative result of treatment on the UPDRS mental subscale (Part I). Given both beneficial and detrimental treatment effects resulting in positive and negative GTE values among outcomes in all treatment/placebo comparisons, the GST leads to the interpretation of no significant global difference between the treatment and the placebo.

INTEGRATING STATISTICAL RESULTS WITH CLINICAL NEUROLOGY

Recently, research has focused on disease modifying treatments that could slow, stop, or reverse disease progression. The identification of such PD therapies is a major unmet need for PD patients and a major goal of neurotherapeutics. Since PD disability is multidimensional, it

is difficult to identify a single most important outcome as the primary outcome to summarize PD disability for the identification of disease modifying treatments. Many multivariate statistical techniques are not designed to test a treatment's global effect. They either lose statistical power (like the QE2 example) or have high power to claim treatment difference if one outcome shows a strong treatment effect (like RAMP example). The GST utilizing GTE is an innovative method to compare treatments based on a treatment's multidimensional performance and provide a single test for global interpretations on whether a new treatment should be advocated.

The QE2 study demonstrated that GST had a higher power when the treatment had a consistent improvement over all outcomes. On the other hand, RAMP study showed that both beneficial and detrimental treatment effects were seen from different outcomes. The GST detected this inconsistent pattern and lowered its power. The single *P*-value from the mental outcome was sufficient, using the Bonferroni adjustment, to conclude that treatment was worse than placebo in the combined treatment group versus placebo comparison. It is worthwhile noting that GST gave a consistent conclusion in the RAMP study as reported by RAMP investigators after an extensive review of the treatment. Among all multivariate techniques considered, MANOVA or the Hotelling's T^2 test was never the most powerful test.

When the objective is to test whether a treatment can improve at least one outcome and it is not intended to address a treatment's global performance, the Bonferroni adjustment and its extensions provide an interpretable answer. However, if investigators need to assess a treatment with multidimensional disease benefit as the primary analysis (as would be important in a long-term trial of a treatment designed to halt or slow the decline of PD) hypothesizing success on one outcome may undermine the analysis. Furthermore, in "neuroprotective" studies, the treatment premise is protection against the disease in all its manifestation as opposed to some symptomatic treatments that may be designed to treat a core constellation of signs (parkinsonism, dyskinesia, cognition, etc.).

The steps used to apply GST in data analysis are straightforward: (1) all observations from both groups are combined and each outcome is ranked separately. For each outcome, the largest rank is given to the best observation of this outcome; (2) each patient's ranks across all outcomes are summed; (3) a two-sample *t* test is performed on these summed ranks with variance adjustment using the method of Huang et al.¹⁵

The steps to compute sample size in study design are: (1) the clinically meaningful difference is determined for each outcome; (2) these differences are converted into GTEs¹⁶ and their average is computed as the final GTE for the GST; (3) the desired significance level and power are chosen; (4) published formula or S+ code¹⁶ can be used to compute sample size.

If the investigator has strong evidence that some outcomes may be more affected by the treatment than others, outcomes may also be weighted to take this prior knowledge into account through slight modification of the formulas in Huang et al.¹⁶ However, the gain in complexity of the interpretation of results, and the potential for disagreement on the chosen weights after the study is reported, may make the use of unweighted outcomes in the GST the most straightforward, and easiest to interpret. This approach has been emphasized in the current analysis.

The GST can be applied in a number of settings when the hypothesis involves multiple related measures. For example, it can be used to test whether a new treatment is efficacious in slowing functional decline or improve the quality of life. It can also be used to test whether a new treatment has fewer side effects or is better tolerated. When the null hypothesis is rejected by the GST, one can look at GTEs from its components and directly identify which one actually contributes to this difference without the need of multiple comparison adjustment. This is

because the overall type I error has already been controlled by the GST. With the unique features of GST and the aid of GTE in its interpretation, we expect that GST will find its broad application in clinical research.

Although tests of higher power are often preferred by investigators, the more important issue with regard to GST is the careful selection of the essential outcomes used to address the proposed scientific question in the analysis. The selection of which outcomes to include in the GST should follow the same principle outlined by International Conference on Harmonization statistical guidelines for multiple primary outcomes in clinical trials: “It should be clear whether an impact on any of the variables, some minimum number of them, or all of them, would be considered necessary to achieve the trial objectives”.³⁹ Only measures that are hypothesized to contribute to the targeted overall benefit should be included in this type of analysis. Because the GTE is an average preference of all outcomes included in the GST, the selection must be tailored to include the minimum number of outcomes that are sufficient to address the scientific question. The choice of outcomes, therefore, should be driven by clinicians who are testing a specific hypothesis and are restricting themselves to core issues. One impediment to the adaptation of this method to the evaluation of clinical trials for the purpose of drug registration and approval for marketing has more to do with the selection of the evaluations which go into the GST than the method itself. To make this method interpretable across different clinical trials of different agents, there must be prior agreement by investigators and regulatory authorities as to the best clinical measures of various PD phenomena, i.e., a “Parkinson-specific” GST. The use of the UPDRS over the years, across all PD trials, has accomplished this in a *de facto* manner. Applying GST to the new MDS-UPDRS and its component subscales may hasten acceptance of the new scale as well as providing an improved method for its interpretation.

Acknowledgments

This work is supported by MCRF grant FHA05CRF and NIH/NINDS grants: U01NS043127 and U01NS043128. The views presented in this article do not necessarily reflect those of the Food and Drug Administration. No official support or endorsement of this article by the Food and Drug Administration is intended or should be inferred.

Financial Disclosure: Peng Huang, Consulting and Advisory Board Membership with honoraria: Data and Safety Monitoring Board (DSMB) for National Institutes of Health (NIH), Grant/Research: NIH, Honoraria: NIH, American College of Physicians, Intellectual Property Rights: none, Salary: Johns Hopkins University School of Medicine; Financial disclosure: Robert Woolson, Consulting and Advisory Board Membership with honoraria: 1. FDA Psychopharm Advisory Committee; 2. NIMH ITV Review Committee Member; 3. Data and Safety Monitoring Board (DSMB) for National Institutes of Health (NIH), VA, & Glaxo (ended in early 2008), Grant/Research: NIH, Honoraria: NIH, VA (DSMB), Pro-ED (Mock FDA Panel completed in 2007/2008), Royalties: Text, Statistical Methods for Analysis of Biomedical Data, John Wiley & Sons, Inc., Salary: Medical University of South Carolina; Financial disclosure: Christopher G. Goetz, MD, Consulting and Advisory Board Membership with honoraria: Allergan, Biogen, Boehringer-Ingelheim, Ceregene, EMD Pharmaceuticals, Embryon, Impax Pharmaceuticals, I3 Research, Juvantia Pharmaceuticals, Kiowa Pharmaceuticals, GlaxoSmith Kline, Merck KgaA, Merck and Co, Neurim Pharmaceuticals, Novartis Pharmaceuticals, Ovation Pharmaceuticals, Oxford Biomedica, Schering-Plough, Solstice Neurosciences, Solvay Pharmaceuticals, Synergy/Intec, Teva Pharmaceuticals, Grants/Research: Funding from NIH, Michael J. Fox Foundation, Kinetics Foundation, and directs the Rush Parkinson's Disease Research Center that receives support from the Parkinson's Disease Foundation, Honoraria: Movement Disorder Society, Northwestern University, American Academy of Neurology, Robert Wood Johnson Medical School, Royalties: Oxford University Press, Elsevier Publishers, Salary: Rush University Medical Center; Financial disclosure: Barbara C. Tilley, Consulting and Advisory Board Membership with honoraria: Data and Safety Monitoring Committees for NIH, and UAB; Advisory Committee for MacroGenics, Grant/Research: NIH, Duke Endowment, Honoraria: NIH, Evanston Health, Intellectual Property Rights: none, Salary: Medical University of South Carolina; Financial disclosure: Douglas Kerr, None of the financial disclosures below were related to or influence in any way the submitted manuscript, Consulting and Advisory Board Membership with honoraria: Nerveda Inc. (Consultant), California Stem Cells Inc. (consultant), Grant/Research: Nerveda Inc., Intellectual Property Rights: Nerveda Inc., Ownership interests: equity stock in Nerveda Inc. (less than \$5,000 value), Royalties: none, Salary: Johns Hopkins University School of Medicine; Financial disclosure: Yuko Y. Palesch, Consulting and Advisory Board Membership with honoraria: None, Grant/Research: NINDS U01 grants, Honoraria: NIH DSMBs, review groups, Salary: Medical University of South Carolina; Financial disclosure: Jordan Elm, Consulting and Advisory Board Membership with honoraria: none, Grant/Research: NIH, Honoraria:

none, Intellectual Property Rights: none, Salary: Medical University of South Carolina; Financial disclosure: Bernard Ravina, Grant support and/or consulting from the following: NIH, Department of Defense, Acadia, Envivo, Novartis, Vernalis, Link, Medivation, Boehringer Ingelheim, Teva, Edison Pharmaceuticals, Michael J. Fox Foundation, Upsher-Smith, Salary: University of Rochester; Financial disclosure: Kenneth J Bergmann, Consulting and Advisory Board Membership with honoraria: None, Grant/Research: NIH, Kyowa Pharmaceutical, Inc., Eisai Pharma, Honoraria: South Carolina Aging Research Conference, 2008: Gene - Environment Interaction in the Pathogenesis of PD, Intellectual Property Rights: None, Ownership interests: None, Royalties: None, Salary: Medical University of South Carolina (to 9/30/2008); Food and Drug Administration (from 10/12/2008 -current); Financial disclosure: Karl Kiebertz, Grant support: Amarin, Boehringer-Ingelheim, Medivation, Neurosearch, NIH, FDA, Consultant: Abbott, Antipodean, Biogen Idec, Ceregene, Eisai, FoldRx, Impax, Ipsen, Lilly, Lundbeck, Merck-Serono, Merz, Novartis, Orion, Prestwick, Schering-Plough, Schwarz, Solvay, Teva, UCB Pharma, Vernalis, FDA, NIH, Legal Consulting: Pfizer, Welding Rod Litigation Defendants. We thank the PSG Steering Committees and the site investigators and coordinators of the QE2 trial and RAMP trials for providing the data for re-analysis. We thank many of our colleagues for helpful discussions and comments.

Author Roles: Research project: Peng Huang, Christopher G. Goetz, Barbara Tilley, Yuko Palesch, Jordan Elm, Bernard Ravina, Kenneth J. Bergmann, Karl Kiebertz; Statistical Analysis: Peng Huang; Manuscript: Peng Huang, Christopher G. Goetz, Robert Woolson, Barbara Tilley, Douglas Kerr, Yuko Palesch, Jordan Elm, Bernard Ravina, MD, Kenneth J. Bergmann, Karl Kiebertz, The Parkinson Study Group.

REFERENCES

1. Fahn, S.; Elton, RL. The unified Parkinson's disease rating scale. In: Fahn, S.; Marsden, CD.; Calne, DB.; Goldstein, M., editors. Recent developments in Parkinson's disease, Vol. 2. Florham Park: Macmillan Healthcare Information; 1987. p. 153-163.p. 293-304.members of the UPDRS Development Committee
2. Goetz CG, Fahn S, Martinez-Martin P, et al. Movement disorder society-sponsored revision of the unified Parkinsons disease rating scale (MDS-UPDRS): process, format, and clinimetric testing plan. *Mov Disord* 2007;22:41–47. [PubMed: 17115387]
3. O'Brien PC. Procedures for comparing samples with multiple endpoints. *Biometrics* 1984;40:1079–1087. [PubMed: 6534410]
4. Pocock SJ, Geller NL, Tsiatis AA. The analysis of multiple endpoints in clinical trials. *Biometrics* 1987;43:487–498. [PubMed: 3663814]
5. Lefkopoulou M, Moore D, Ryan L. The analysis of multiple correlated binary outcomes: application to rodent teratology experiments. *J Am Stat Assoc* 1989;84:810–815.
6. Tang D-I, Gnecco C, Geller NL. An approximate likelihood ratio test for a normal mean vector with nonnegative components with application to clinical trials. *Biometrika* 1989;76:751–754.
7. Tang D, Gnecco C, G'er NL. Design of group sequential clinical trials with multiple endpoints. *J Am Stat Assoc* 1989;84:776–779.
8. Lehman W, Wassmer G, Reitmeir P. Procedures for two-sample comparisons with multiple endpoints controlling the experiment-wise error rate. *Biometrics* 1991;47:511–521. [PubMed: 1912258]
9. Lefkopoulou M, Ryan L. Global tests for multiple binary outcomes. *Biometrics* 1993;49:975–988. [PubMed: 8117908]
10. Tang D, Geller NL, Pocock SJ. On the design and analysis of randomized clinical trials with multiple endpoints. *Biometrics* 1993;49:23–30. [PubMed: 8513104]
11. Legler JM, Lefkopoulou M, Ryan LM. Efficiency and power of tests for multiple binary outcomes. *J Am Stat Assoc* 1995;90:680–693.
12. Karrison TG, O'Brien PC. A rank-sum-type test for paired data with multiple endpoints. *J Applied Stat* 2004;31:229–238.
13. Tang D, Geller NL. Closed testing procedures for group sequential clinical trials with multiple endpoints. *Biometrics* 1999;55:1188–1192. [PubMed: 11315066]
14. Roy J, Lin X. Latent variable models for longitudinal data with multiple continuous outcomes. *Biometrics* 2000;56:1047–1054. [PubMed: 11129460]
15. Huang P, Tilley B, Woolson R, Lipsitz S. Adjusting O'Brien's test to control type I error for the generalized nonparametric Behrens-Fisher problem. *Biometrics* 2005;61:532–539. [PubMed: 16011701]

16. Huang P, Woolson RF, O'Brien PC. A rank-based sample size method for multiple outcomes in clinical trials. *Stat Med* 2008;27:3084–3104. [PubMed: 18189338]
17. Huang P, Woolson RF, Granholm A. The use of a global statistical approach for the design and data analysis of clinical trials with multiple primary outcomes. *Experimental stroke* 2009;1:100–109.
18. Tilley BC, Marler J, Geller NL, et al. Use of a global test for multiple outcomes in stroke trials with application to the national institute of neurological disorders and stroke t-PA stroke trial. *Stroke* 1996;27:2136–2142. [PubMed: 8898828]
19. Kwiatkowski T, Libman R, Frankel M, et al. NINDS rt-PA Stroke Study Group. Effects of tissue plasminogen activator for acute ischemic stroke at one year. *N Engl J Med* 1999;340:1781–1787. [PubMed: 10362821]
20. Kaufman KD, Olsen EA, Whiting D, Savin R, De Villez R, Bergfeld W. Finasteride in the treatment of men with androgenetic alopecia. *J Am Acad Dermatol* 1998;39:578–589. [PubMed: 9777765]
21. Li DK, Zhao GJ, Paty DW. Randomized controlled trial of interferon-beta-1a in secondary progressive MS: MRI results. *Neurology* 2001;56:1505–1513. [PubMed: 11402107]
22. Shames RS, Heilbron DC, Janson SL, Kishiyama JL, Au DS, Adelman DC. Clinical differences among women with and without self-reported perimenstrual asthma. *Ann Allergy Asthma Immunol* 1998;81:65–72. [PubMed: 9690575]
23. Tilley BC, Pillemer SR, Heyse SP, Li S, Clegg DO, Alarcon GS. Global statistical tests for comparing multiple outcomes in rheumatoid arthritis trials. *Arthritis Rheum* 1999;42:1879–1888. [PubMed: 10513802]
24. Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat* 1979;6:65–70.
25. Simes RJ. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 1986;73:751–754.
26. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 1988;75:800–802.
27. Hommel G. A stagewise multiple test procedure based on a modified Bonferroni test. *Biometrika* 1988;75:383–386.
28. Hommel G. A comparison of two modified Bonferroni procedures. *Biometrika* 1989;76:624–625.
29. Lehman W, Wassmer G, Reitmeir P. Procedures for two-sample comparisons with multiple endpoints controlling the experiment-wise error rate. *Biometrics* 1991;47:511–521. [PubMed: 1912258]
30. Matser E, Kessels A, Lezak M, Jordan B, Troost J. Neuropsychological impairment in amateur soccer players. *JAMA* 1999;282:971–973. [PubMed: 10485683]
31. Wetter TC, Stiasny K, Winkelmann J, et al. A randomized controlled study of pergolide in patients with restless legs syndrome. *Neurology* 1999;52:944–950. [PubMed: 10102410]
32. van Kleef M, Barendse GAM, Kessels A, Voets HM, Weber WEJ, de Lange S. Randomized trial of radiofrequency lumbar facet denervation for chronic low back pain. *Spine* 1999;24:1937–1942. [PubMed: 10515020]
33. Goodkin DE, Rudick R, Medendorp SV, et al. Low-dose (7.5 mg) oral methotrexate reduces the rate of progression in chronic progressive multiple sclerosis. *Ann Neurol* 1995;37:30–40. [PubMed: 7818255]
34. Poole CJ, Earl HM, Hiller L, et al. for the NEAT Investigators and the SCTBG. Epirubicin and cyclophosphamide, methotrexate, and fluorouracil as adjuvant therapy for early breast cancer. *N Engl J Med* 2006;355:1851–1862. [PubMed: 17079759]
35. Acion L, Peterson J, Temple S, Arndt S. Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects. *Stat Med* 2006;25:591–602. [PubMed: 16143965]
36. Wolfe DA, Hogg RV. On constructing statistics and reporting data. *Am Stat* 1971;25:27–30.
37. Shults CW, Oakes D, Kieburtz K, et al. Parkinson Study Group. Effects of coenzyme Q₁₀ in early Parkinson disease: evidence of slowing of the functional decline. *Arch Neurol* 2002;59:1541–1550. [PubMed: 12374491]
38. Parkinson Study Group. A multicenter randomized controlled trial of remacemide hydrochloride as monotherapy for PD. *Neurology* 2000;54:1583–1588. [PubMed: 10762497]

39. ICH Expert Working Group. Guideline E9: Statistical Principles for Clinical Trials. International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use. Available at. <http://www.ich.org/cache/compo/276-254-1.html>.

TABLE 1**Multivariate analytical procedures for multiple clinical outcomes**

Method: Choose one outcome as the primary, other outcomes are analyzed as secondary.

Aim: To test treatment difference on this single primary outcome.

Representative examples: Univariate tests such as Wilcoxon test, t-test, etc.

Advantages

- Simple in sample size computation and data analysis.
- The interpretation of the data is straightforward if this primary outcome is clearly the most important outcome.

Limitations

- When multiple outcomes are equally important to assess a treatment's benefit, it can lose power when treatment shows benefit on majority of secondary outcomes.
- Results can be difficult to interpret if the primary analysis and the secondary analysis have quite different conclusions.

Method: Form a linear combination of all outcomes.

Aim: To test treatment difference on this composite outcome.

Representative examples: Weighted average, principal component analysis.

Advantages

- Potential high statistical power.
- Easy to interpret if all outcomes are measuring similar quantities and are of the same type (e.g. all continuous or all non-continuous).

Limitations

- When outcomes measured in different scales (e.g., continuous and non-continuous), it can be difficult to construct a composite score. For example, to combine total UPDRS and the delay of levodopa initiation could lead to practical difficulties for statistical analysis or forced categorizations of the UPDRS and time to levodopa initiation that can be arbitrary.
- Interpretation of the composite score can be difficult.
- Statistical conclusion can be altered when a non-linear transformation is applied to the data

Method: Multiple tests with adjustment to the overall significance level.

Aim: To test whether there is any treatment difference on any single outcome.

Representative examples: Bonferroni adjustment, Simes test, Hochberg procedure.

Advantages

- Appropriate to test whether a treatment has any effect on any of these outcomes.
- High statistical power if a treatment has a strong effect on one outcome.

Limitations

- Lack a global assessment of a treatment's benefit on multiple outcomes, especially when treatment demonstrates both beneficial and detrimental effects on different outcomes.
- Potential low power when outcomes are highly correlated.

Method: Omnidirectional test.

Aim: To test whether the treatment changes the joint distribution of multiple outcomes.

Representative examples: Hotelling's T^2 test, MANOVA, Wald test, χ^2 test.

Advantages

- Appropriate to test whether a treatment has any type of effect on any of these outcomes.
- High statistical power if a treatment has a strong effect on one outcome.

Limitations

- Lack a global assessment of a treatment's benefit on multiple outcomes, especially when treatment demonstrates both beneficial and detrimental effects on different outcomes.
- Strong mathematical assumptions on the joint outcome probability distributions.

Method: Hierarchical model.

Aim: To test whether the treatment improves all outcomes.

Representative examples: Latent model, Bayes model, conditional independent model.

Advantages

- Flexible in modeling all types of outcomes and analyzing them together.
- Good mathematical properties.
- Easy to design a clinical trial to test the latent parameter or the hyperparameter.

Limitations

- Treatment benefit is expressed as the improvement of all outcomes that can be violated in practice.
- Difficult to verify the validity of the model assumption, particularly the assumption how multiple outcomes are associated with each others.
- Difficult for clinicians to understand the latent parameter or hyperparameter and difficult to translate into clinical practice in terms of the amount of benefit a patient could receive.
- Strong mathematical assumptions on probability distributions that are difficult to verify.

Method: Global statistical test (GST)

Aim: To test whether the treatment has a global benefit based on multiple pivotal outcomes.

Representative examples: GST using ranks, GST using least squares principle.

Advantages

- Useful to test a treatment's global benefit across different outcomes and determine whether a treatment is preferred to use.
- High statistical power when treatment shows benefit on majority of outcomes
- Avoids misinterpretation of data if some outcomes show improvement but at the price of serious decline in other key target outcomes.

Limitations

- Lose power if the objective is to detect any type of treatment difference when both beneficial and detrimental effects are seen from different outcomes.

TABLE 2

QE2 trial: change from baseline to last visit

Comparison	Outcome	GTE	<i>P</i> value*
Combined	<u>Univariate Test</u>		
CoQ ₁₀	Motor	0.1191	0.4664
Group	Mental	0.2158	0.1543
(<i>n</i> = 64)	ADL	0.3154	0.0514
Versus	Total UPDRS	0.2490	0.1263
Placebo Group	<u>Multivariate test^a</u>		
(<i>n</i> = 16)	Hotelling's T ²	NA	0.1264
	Bonferroni	NA	0.1542
	GST ^b	0.2168	0.0956
	<u>Univariate Test</u>		
1200 mg/d	Motor	0.2065	0.2840
CoQ ₁₀	Mental	0.2418	0.1782
Group	ADL	0.4565	0.0163
(<i>n</i> = 23)	Total UPDRS	0.3696	0.0538
Versus	<u>Multivariate test^a</u>		
Placebo Group	Hotelling's T ²	NA	0.0940
(<i>n</i> = 16)	Bonferroni	NA	0.0489
	GST ^b	0.3016	0.0389
	<u>Univariate Test</u>		
600 mg/d	Motor	0.1000	0.6212
CoQ ₁₀	Mental	0.2750	0.1387
Group	ADL	0.0781	0.7009
(<i>n</i> = 20)	Total UPDRS	0.1625	0.4166
Versus	<u>Multivariate test^a</u>		
Placebo Group	Hotelling's T ²	NA	0.3935
(<i>n</i> = 16)	Bonferroni	NA	0.4161
	GST ^b	0.1510	0.3255
	<u>Univariate Test</u>		
300 mg/d CoQ ₁₀	Motor	0.0417	0.8418
Group (<i>n</i> = 21)	Mental	0.1310	0.4864
Versus	ADL	0.3869	0.0461
Placebo Group	Total UPDRS	0.1994	0.3109
(<i>n</i> = 16)	<u>Multivariate test^a</u>		
	Hotelling's T ²	NA	0.2445
	Bonferroni	NA	0.1383
	GST ^b	0.1865	0.2020

* *P*-values for univariate test were obtained from the Wilcoxon rank test. *P*-value of Bonferroni adjustment was obtained by multiplying the smallest *P*-value from Motor, Mental, and ADL by three (the number of outcomes included in the multivariate test).

^a Outcomes included in the multivariate test were: Motor, Mental, and ADL. MANOVA is not listed because it gave the same results as Hotelling's T^2 test.

^b The GST is from Huang et al.¹⁵

TABLE 3

RAMP trial: comparing change from baseline to last visit

Comparison	Outcome	GTE	<i>P</i> value*
	<u>Univariate Test</u>		
Combined	Mental	−0.2355	0.0041
Treatment	Motor + ADL	−0.0389	0.6811
(<i>n</i> = 150)	BDI	−0.0552	0.5292
Versus	MMSE	0.0421	0.6404
Placebo	<u>Multivariate Test</u> **		
(<i>n</i> = 50)	Bonferroni	NA	0.0164
	MANOVA	NA	0.4273
	GST	−0.0719	0.1456
	<u>Univariate Test</u>		
Treatment	Mental	−0.2498	0.0144
600 mg	Motor + ADL	0.0151	0.8975
(<i>n</i> = 53)	BDI	−0.1532	0.1476
Versus	MMSE	−0.0075	0.9476
Placebo	<u>Multivariate Test</u> **		
(<i>n</i> = 50)	Bonferroni	NA	0.0576
	MANOVA	NA	0.1057
	GST	−0.0989	0.0809
	<u>Univariate Test</u>		
Treatment	Mental	−0.2435	0.0170
300 mg	Motor + ADL	−0.0178	0.8832
(<i>n</i> = 46)	BDI	−0.0513	0.6398
Versus	MMSE	0.0422	0.7108
Placebo	<u>Multivariate Test</u> **		
(<i>n</i> = 50)	Bonferroni	NA	0.0680
	MANOVA	NA	0.2630
	GST	−0.0676	0.2440
	<u>Univariate Test</u>		
Treatment	Mental	−0.2133	0.0421
150 mg	Motor + ADL	−0.1141	0.3241
(<i>n</i> = 51)	BDI	0.0431	0.6862
Versus	MMSE	0.0937	0.3974
Placebo	<u>Multivariate Test</u> **		
(<i>n</i> = 50)	Bonferroni	NA	0.1684
	MANOVA	NA	0.6164
	GST	−0.0476	0.4443

* *P*-values for univariate outcomes were obtained from Wilcoxon rank test. *P*-value of Bonferroni adjustment was obtained by multiplying the smallest *P*-value from Mental, Motor, ADL, BDI, SEADL, and MMSE by 4 (the number of outcomes included in the multivariate test).

** Outcomes included in the multivariate test were: Motor, Mental, ADL, BDI, SEADL, and MMSE. The GST is from Huang et al.¹⁵