A rank-based test for clinical trials with multivariate longitudinal outcomes

Xiaoming Xu

Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina

email: xiaoming.xu197@duke.edu

Sheng Luo

Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina

email: sheng.luo@duke.edu

Abstract

Clinical trials of many chronic diseases such as Parkinson's disease often collect multiple health outcomes to monitor the disease severity and progression. It is of scientific interest to test whether the experimental treatment has an overall efficacy on the multiple outcomes across time, as compared to placebo or an active control. To compare the multivariate longitudinal outcomes between two groups, the rank-sum test[1] and the variance-adjusted rank-sum test[2] can be used to test the treatment efficacy. But these two rank-based tests, by utilizing only the change from baseline to the last time point, do not fully take advantage of the multivariate longitudinal outcome data, and thus may not objectively evaluate the global treatment effect over the entire therapeutic period. In this paper, we develop rank-based test procedures to detect global treatment efficacy in clinical trials with multiple longitudinal outcomes. We first conduct an interaction test to determine whether treatment effect varies over time, and then propose a longitudinal rank-sum test to assess the main treatment effect either with or without the interaction. Asymptotic properties of the proposed test procedures are derived and thoroughly examined. Simulation studies under various scenarios are performed. The test statistic is motivated by and applied to a recently-completed randomized controlled trial of Parkinson's disease.

Key words: Nonparametrics; Global test; Parkinson's disease; rank-sum-type test; U-Statistics

1 Introduction

Parkinson's disease (PD) is a chronic progressive neurodegenerative disorder that affects about 1% of people older than 60 years in the United States alone.[3] PD causes impairment in multiple domains (e.g., motor, cognitive, and behavioral). The disease progresses heterogeneously in time and across domains and individuals: decline may be observed in some, but not all health outcomes at any given time interval and the trajectory of progression may vary between different domains, both within and across PD patients. Therefore, no single health outcome reliably reflects the full spectrum of the disease severity and progression. Randomized clinical trials (RCTs) of PD repeatedly collect multiple health outcomes to obtain an overview of disease progression of PD patients.[4–6] For example, Azilect study is a multi-center, placebo-controlled, Phase 3 study to evaluate the efficacy of rasagiline in Japanese patients with early PD.[6] The patients with a diagnosis of PD within 5 years were randomized 1:1 to receive rasagiline (1 mg/day) or placebo for up to 26 weeks. The primary endpoint is the Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS) with four parts (Part I to IV) and it was measured at baseline, weeks 6, 10, 14, 20, and 26. The multiple outcomes being considered were changes from baseline to each visit in MDS-UPDRS Part I sum score, Part II sum score, and Part III sum score.

To analyze the multivariate longitudinal outcome data from RCTs comparing an experimental treatment with placebo or an active control, an appropriate statistical method needs to fully utilize the multivariate longitudinal outcome data and accounts for three sources of correlation: (1) intra-source (same outcome at different visits), (2) inter-source (different outcomes at the same visit), and (3) cross correlation (different outcomes at different visits).[7] Additionally, the scientific interest of many RCTs is directional, i.e., whether the experimental treatment is better than the placebo or active control based on all outcomes considered. Hence, the research goal is to evaluate the global treatment efficacy across multiple longitudinal outcomes.

Current major analysis methods for multivariate longitudinal data is to reduce the longitudinal data to cross-sectional data by computing the change from baseline to the last observation and

then adopt some global procedures to test the treatment efficacy. Broadly speaking, the global procedures can be classified into two categories: parametric procedure and nonparametric procedure. The parametric procedure, under the assumption of multivariate normality, includes likelihood ratio tests[8–12] and t-statistics type tests.[1,13–15] The distribution assumptions associated with the parametric procedure, however, often restrict its use in practice.

The nonparametric rank-based procedure, on the other hand, is distribution-free and widely used in clinical research. To list a few, there are the statistical tests given by Akritas and Brunner,[16] Brunner et al.,[17] Brunner et al.,[18] Roy et al.,[19] Gunawardana and Konietschke,[20] Dobler et al.[21] and so on. Specifically, Brunner et al.[17] developed Wald-type and ANOVA-type tests to assess whether two treatment groups differ, in which the multiple outcomes are considered simultaneously. Roy et al.[19] considered nonparametric methods for two-sample problems in which each subjects may have an individual number of correlated replicates. O'Brien[1] proposed a rank-sum-type test to assess whether outcome measures from the treatment group are better than those from the control group. Although this test is robust and efficient in testing treatment effect across the multiple outcomes, it has an inflated type I error rate when the two groups being compared have different variances. Huang et al.[2] examined the theoretical properties of this test and proposed a variance-adjusted rank-sum test that controls the type I error rate. Moreover, Liu et al.,[22] by taking the maximum of the absolute value of individual rank-sum statistics, proposed a rank-max-type test statistic. When the group differences in all outcomes lie in the opposite direction (i.e., treatment may have efficacy in some outcomes but negative effect in the other outcomes, as compared to placebo), the test of Liu et al.,[22] compared with rank-sum tests of O'Brien[1] and Huang et al.,[2] maintains satisfactory power in detecting the group-differences across all outcomes. Further, Zhang et al.[23] developed a cluster-adjusted rank-based test, in which the outcomes are divided into several clusters such that treatment effects are expected to be more similar within clusters than between clusters. The cluster-adjusted rank test has the strength of both the rank-sum-type tests and the rank-max-type test by accumulating the evidence within each cluster with a rank-sum-type test and combining the cluster-level evidence using the max test.

However, all these rank-based tests are only applicable to the cross-sectional change data, do not

fully take advantage of the multivariate longitudinal outcome data. They are not able to track how the relative treatment efficacy changes over time (which might be of scientific interest) and thus may not objectively evaluate the treatment effect throughout the whole treatment period. Furthermore, while the rank-sum tests of O'Brien[1] and Huang et al.[2] can evaluate whether treatment is effective regarding all the outcomes considered, and provide a directional conclusion, other tests such as the Wald-type and ANOVA-type tests of Brunner et al.,[17] Brunner et al.,[18] Roy et al.,[19] Dobler et al.,[21] are not directional and are unable to evaluate the treatment efficacy even after rejecting the null hypothesis of no difference between two groups.

In addition, rank-based nonparametric procedures were developed in repeated measures designs. The closely related works include Konietschke et al.,[24] Zhuang et al.[25], Umlauft et al.,[26] Rubarth et al.,[27] Rubarth et al.,[28] and so on. In these works, only one longitudinal outcome was considered and thus only the intra-source correlation of this outcome was accounted for. Thus, the comparison between two groups based on a single outcome variable cannot fully reflect the overall treatment effects when all outcomes reflecting different disease aspects need to be considered simultaneously. Further, most of these proposed tests are not directional, because they are Wald-type and ANOVA-type tests to assess whether two treatment groups differ.

In this paper, we develop a nonparametric global test procedure to detect treatment efficacy in clinical trials with multiple longitudinal outcomes. We first conduct an interaction test to see if treatment effect varies over time, and then propose a longitudinal rank-sum test to assess the main treatment effect either with or without the interaction. Compared with the aforementioned rank-based tests, the proposed test procedure not only fully utilizes the longitudinal data in multiple outcomes, but also provides a directional conclusion of main treatment efficacy in the whole therapeutic period. Because the longitudinal rank-sum test extends the rank-sum test in Huang et al.[2] (referred to as Huang's test) by incorporating the rank-sum test statistic from each time point, we compare it with Huang's test in terms of main treatment effect in the simulation study and real data analysis.

The rest of the paper is organized as follows. Section 2 gives some preliminary results, develops the interaction test and longitudinal rank-sum test, and examines their theoretical properties. In

Section 3, extensive simulations are conducted to investigate their performance in type I error and power. In Section 4, we apply the proposed test to the Azilect study of Parkinson's disease. Some discussions and future work are given in section 5. All technical details and additional results are provided in the Supplemental materials.

2 Methods

2.1 Notations and Preliminary Results

We consider two groups of subjects, treatment vs. control, who are followed in a longitudinal study with T assessment times. At each time point, a total of K outcomes are measured on each subject. Let $\boldsymbol{X} = (\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_T^\top)^\top$ and $\boldsymbol{Y} = (\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_T^\top)^\top$ represent the multiple outcome variables over time for the control group and treatment group, respectively, where $\boldsymbol{X}_t = (X_{t1}, \cdots, X_{tK})^\top$ and $\boldsymbol{Y}_t = (Y_{t1}, \cdots, Y_{tK})^\top$, $t = 1, \cdots, T$. Without the loss of generality, we let larger values represent better clinical results for all outcomes. Let X_{tk} and Y_{tk} have marginal distributions F_{tk} and G_{tk}, respectively. Further, for outcome k and at time t, we define $\theta_{tk} = P(X_{tk} < Y_{tk}) - P(X_{tk} > Y_{tk})$. According to Brunner et al.[29], the parameter θ_{tk} is called the relative effect of the Y-group with respect to the X-group for the kth outcome at the time point t. Denote $\theta_t = \frac{1}{K}\sum_{k=1}^{K}\theta_{tk}$, which is a measure of the relative effect of treatment group with respect to control group across all outcomes at time t.

Before formally setting up the null hypotheses and the associated test statistics, we introduce some notation and preliminary results that are essential to the inferential procedures. Let x_{itk} $(i = 1, \cdots, m)$ be the response of outcome k from subject i in the control group at time t. Similarly, we denote y_{jtk} $(j = 1, \cdots, n)$ to be the response in the treatment group. Let $N = m + n$ be the total number of subjects in two groups. For outcome k at time t, we combine all observations in two groups $x_{1tk}, \cdots, x_{mtk}$ and $y_{1tk}, \cdots, y_{ntk}$ and rank them, with larger values obtaining higher ranks. Denote the mid-ranks of x_{itk} and y_{jtk} by R_{xitk} and R_{yjtk}, respectively, which is either the regular rank when there is no tie on the observations or the average rank of those tied observations.

Further, we define:

$$\bar{R}_{x\cdot tk} = \frac{1}{m}\sum_{i=1}^{m} R_{xitk}, \quad \bar{R}_{x\cdot t\cdot} = \frac{1}{mK}\sum_{i=1}^{m}\sum_{k=1}^{K} R_{xitk},$$

and

$$\bar{R}_{y\cdot tk} = \frac{1}{n}\sum_{j=1}^{n} R_{yjtk}, \quad \bar{R}_{y\cdot t\cdot} = \frac{1}{nK}\sum_{j=1}^{n}\sum_{k=1}^{K} R_{yjtk},$$

for X-sample and Y-sample, respectively. Notice that $\hat{\theta}_{tk} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}[I(x_{itk} < y_{jtk}) - I(x_{itk} > y_{jtk})] = \frac{2}{N}(\bar{R}_{y\cdot tk} - \bar{R}_{x\cdot tk})$ is a consistent estimator of θ_{tk} (see the detailed derivations in Supplemental materials A). It follows that $\hat{\theta}_t = \frac{2}{N}(\bar{R}_{y\cdot t\cdot} - \bar{R}_{x\cdot t\cdot})$ is also a consistent estimator of $\theta_t = \frac{1}{K}\sum_{k=1}^{K}\theta_{tk}$. Hence, larger values of $(\bar{R}_{y\cdot tk} - \bar{R}_{x\cdot tk})$ and/or $(\bar{R}_{y\cdot t\cdot} - \bar{R}_{x\cdot t\cdot})$ gives evidence that the treatment has better effect as compared with the control.

Next, let $\boldsymbol{R} = (\bar{R}_{y\cdot 1\cdot} - \bar{R}_{x\cdot 1\cdot}, \cdots, \bar{R}_{y\cdot T\cdot} - \bar{R}_{x\cdot T\cdot})^\top$. It is worth noting that the rank difference vector $\boldsymbol{R}$ is a two sample U-statistic and thus it follows a multivariate normal distribution asymptotically. We establish the asymptotic joint distribution of the rank difference vector $\boldsymbol{R}$ as follows.

Theorem 1. *The rank difference vector $\frac{1}{\sqrt{N}}\boldsymbol{R}$ asymptotically follows a multivariate normal distribution with mean $\frac{\sqrt{N}}{2}(\theta_1, \cdots, \theta_T)^\top$ and covariance matrix $\boldsymbol{\Sigma}_{T\times T}$ as $\min(m,n) \to \infty$ and $\frac{m}{n} \to \lambda < \infty$, where $\{\boldsymbol{\Sigma}\}_{t_1t_2}$ can be approximated by:*

$$\{\boldsymbol{\Sigma}\}_{t_1t_2} = \frac{1}{K^2}\sum_{k_1=1}^{K}\sum_{k_2=1}^{K}\left[\left(1+\frac{1}{\lambda}\right)c_{t_1k_1,t_2k_2} + (1+\lambda)\,d_{t_1k_1,t_2k_2}\right], \tag{2.1}$$

where $c_{t_1k_1,t_2k_2} = cov(G_{t_1k_1}(X_{t_1k_1}), G_{t_2k_2}(X_{t_2k_2}))$ and $d_{t_1k_1,t_2k_2} = cov(F_{t_1k_1}(Y_{t_1k_1}), F_{t_2k_2}(Y_{t_2k_2}))$.

Please refer to Supplemental materials A for the detailed proof. From the expression of $\{\boldsymbol{\Sigma}\}_{t_1t_2}$ in Equation (2.1), the three sources of correlation for the rank differences $(\bar{R}_{y\cdot tk} - \bar{R}_{x\cdot tk})$ can be explained by the covariance matrix $\boldsymbol{\Sigma}$. Specifically, from Equation (2.1),

$$\left(1+\frac{1}{\lambda}\right)c_{t_1k,t_2k} + (1+\lambda)\,d_{t_1k,t_2k},\ k = 1, \cdots, K,\ t_1 \neq t_2,$$

$$\left(1+\frac{1}{\lambda}\right)c_{tk_1,tk_2} + (1+\lambda)\,d_{tk_1,tk_2},\ t = 1, \cdots, T,\ k_1 \neq k_2,$$

and

$$\left(1+\frac{1}{\lambda}\right) c_{t_1k_1,t_2k_2} + (1+\lambda)\, d_{t_1k_1,t_2k_2},\ t_1 \neq t_2,\ k_1 \neq k_2$$

account for the intra-source correlation, inter-source correlation, and cross correlation, respectively. The following result gives the moment estimates of $c_{t_1k_1,t_2k_2}$ and $d_{t_1k_1,t_2k_2}$, and the proof is given in Supplemental materials B.

Theorem 2. *Under the conditions in Theorem 1, the consistent estimators of* $c_{t_1k_1,t_2k_2}$ *and* $d_{t_1k_1,t_2k_2}$ *are given by*

$$\hat{c}_{t_1k_1,t_2k_2} = \frac{1}{m}\sum_{i=1}^{m}\left\{\left[\frac{1}{n}\sum_{j=1}^{n} I(y_{jt_1k_1} < x_{it_1k_1}) - \frac{1-\hat{\theta}_{t_1k_1}}{2}\right]\left[\frac{1}{n}\sum_{j=1}^{n} I(y_{jt_2k_2} < x_{it_2k_2}) - \frac{1-\hat{\theta}_{t_2k_2}}{2}\right]\right\}, \tag{2.2}$$

and

$$\hat{d}_{t_1k_1,t_2k_2} = \frac{1}{n}\sum_{j=1}^{n}\left\{\left[\frac{1}{m}\sum_{i=1}^{m} I(x_{it_1k_1} < y_{jt_1k_1}) - \frac{1+\hat{\theta}_{t_1k_1}}{2}\right]\left[\frac{1}{m}\sum_{i=1}^{m} I(x_{it_2k_2} < y_{jt_2k_2}) - \frac{1+\hat{\theta}_{t_2k_2}}{2}\right]\right\}, \tag{2.3}$$

respectively, where $\hat{\theta}_{tk} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}[I(x_{itk} < y_{jtk}) - I(x_{itk} > y_{jtk})]$, *for* $t = 1, 2, \cdots, T$, $k = 1, 2, \cdots, K$.

By plugging the estimates $\hat{c}_{t_1k_1,t_2k_2}$ and $\hat{d}_{t_1k_1,t_2k_2}$, $k_1, k_2 \in \{1, \cdots, K\}$, $t_1, t_2 \in \{1, \cdots, T\}$, we obtain the covariance estimate $\hat{\mathbf{\Sigma}}$. For ease of computation, a computationally efficient way for estimation of $\{\mathbf{\Sigma}\}_{t_1t_2}$ are presented in Supplemental materials C.

2.2 Interaction test

When analyzing longitudinal data in RCTs, the primary objective is to test the significance of the treatment and time interaction term, i.e., the slope differences. In the absence of treatment and time interaction, the treatment effect is constant over time and the outcome profiles of two groups are parallel graphically. Hence, testing whether the interaction is significant is equivalent to testing whether the group profiles are of the same shape. In this field, a closely related work proposed in Zhuang et al.[25] provides a rank-based non-parametric interaction test for longitudinal ordinal

data.

Within this non-parametric framework, we propose the null and alternative hypotheses of interest for testing interaction as follows:

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_T \quad vs \quad H_1 : \text{ not } H_0 \tag{2.4}$$

where $\theta_t = \frac{1}{K}\sum_{k=1}^{K}\theta_{tk} = \frac{1}{K}\sum_{k=1}^{K}[P(X_{tk} < Y_{tk}) - P(X_{tk} > Y_{tk})]$. When the multiple outcomes from the two groups are parallel, we expect that θ_ts would be roughly the same across different time points. Rejection of the null suggests that the interaction is significant, indicating that the averaged relative effect θ_t of the treatment group with respect to the control group changes over time. Let $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_T)^\top$ and define $\boldsymbol{C}$ such that $\{\boldsymbol{C}\}_{s,s} = 1$, $\{\boldsymbol{C}\}_{s,s+1} = -1$ ($s = \{1, \cdots, T-1\}$) and 0 otherwise. Then, testing (2.4) is equivalent to testing $\boldsymbol{C\theta} = \boldsymbol{0}$.

To test the interaction effect, we propose the following wald type test statistic:

$$T_{int} = \frac{1}{N}(\boldsymbol{CR})^\top(\boldsymbol{C\hat{\Sigma}C}^\top)^{-1}(\boldsymbol{CR}). \tag{2.5}$$

Larger values of T_{int} gives evidence that the interaction effect is significant. By Theorem 1, it is easy to see that $\frac{1}{\sqrt{N}}\boldsymbol{CR}$ converges to multivariate normal distribution with mean $\frac{\sqrt{N}}{2}\boldsymbol{C\theta}$ and covariance matrix $\boldsymbol{C\Sigma C}^\top$ asymptotically. Hence, under the null, $(\boldsymbol{C\Sigma C}^\top)^{-\frac{1}{2}}\frac{1}{\sqrt{N}}\boldsymbol{CR} \sim MVN(\boldsymbol{0}, \boldsymbol{I}_{T-1})$, and it follows that the quantity $T^*_{int} = \frac{1}{N}(\boldsymbol{CR})^\top(\boldsymbol{C\Sigma C}^\top)^{-1}(\boldsymbol{CR})$ have χ^2_{T-1} distribution asymptotically. By plugging the consistent covariance estimate $\hat{\boldsymbol{\Sigma}}$, we have that $T_{int} \sim \chi^2_{T-1}$ by Slutsky's theorem under the null. The established χ^2 density of T_{int} allows one to compute the power function of the proposed test under the null and alternative hypothesis, the critical value for a given significance level, and the p-value of the test as well.

Compared with the interaction test in Zhuang et al.[25], the proposed interaction test has several advantages: (1) it applies to different type of data (continuous, ordinal and binary), while in Zhuang et al.[25] only longitudinal ordinal data were considered; (2) it does not require the variance adjustment, which is necessary in the test in Zhuang et al.[25] for tied observations; (3) Zhuang et al.[25] assumes that baseline information is different between the two groups, which may not be the

case in randomized clinical trials due to randomization.

2.3 Global main effect test

2.3.1 Main effect test when an interaction exists

Significant interaction test suggests that the treatment effect relative to control changes over time. Suppose an interaction effect exists, the main treatment effect should be tested by averaging the relative treatment effect across all time points. To test the global treatment efficacy in this scenario, the null and alternative of the main effect is given as follows:

$$H_0 : \bar{\theta} = 0 \quad vs \quad H_1 : \bar{\theta} > 0 \tag{2.6}$$

where $\bar{\theta} = \frac{1}{T}\sum_{t=1}^{T}\theta_t$. For testing (2.6), the test statistics proposed in Huang et al.[2] (referred to as Huang's test), which focused on the treatment effect at a single time point (the last time point), are no longer suitable. In the presence of treatment and time interaction, an appropriate test statistic should account for the treatment efficacy across all the time points. To test the global main treatment effect, we propose the longitudinal rank-sum test (LRST) statistic

$$T_{LRST} = \frac{\bar{R}_{y\cdots} - \bar{R}_{x\cdots}}{\sqrt{\widehat{var}(\bar{R}_{y\cdots} - \bar{R}_{x\cdots})}}, \tag{2.7}$$

where $\bar{R}_{x\cdots} = \frac{1}{mTK}\sum_{i=1}^{m}\sum_{k=1}^{K}\sum_{t=1}^{T} R_{xitk}$, $\bar{R}_{y\cdots} = \frac{1}{nTK}\sum_{j=1}^{n}\sum_{k=1}^{K}\sum_{t=1}^{T} R_{yitk}$ and $\widehat{var}(\bar{R}_{y\cdots} - \bar{R}_{x\cdots})$ is a consistent estimator of the variance of $\bar{R}_{y\cdots} - \bar{R}_{x\cdots}$. While the rank-sum test statistics of O'Brien[1] and Huang et al.[2] are linear combinations of the rank differences across outcomes, the proposed longitudinal rank-sum test statistic T_{LRST} can be viewed as a linear combination of the rank differences $(\bar{R}_{y\cdot tk} - \bar{R}_{x\cdot tk})$ across both outcomes and time points so that the longitudinal data can be fully utilized and treatment effect are evaluated over the whole treatment period.

Let J be a vector of length T with all 1's. From Theorem 1, $T^*_{LRST} = \frac{\bar{R}_{y\cdots} - \bar{R}_{x\cdots}}{\sqrt{var(\bar{R}_{y\cdots} - \bar{R}_{x\cdots})}} = \frac{J^\top \boldsymbol{R}}{\sqrt{J^\top var(\boldsymbol{R}) J}} = \frac{J^\top \boldsymbol{R}}{\sqrt{J^\top N \boldsymbol{\Sigma} J}}$ follows a standard normal distribution asymptotically under the null. By plugging the consistent covariance estimator $\hat{\boldsymbol{\Sigma}}$, we have that the proposed longitudinal rank-sum

test statistic $T_{LRST} = \frac{\bar{R}_{y\cdots} - \bar{R}_{x\cdots}}{\sqrt{\widehat{var}(\bar{R}_{y\cdots} - \bar{R}_{x\cdots})}} = \frac{J^\top \boldsymbol{R}}{\sqrt{J^\top N \hat{\boldsymbol{\Sigma}} J}}$ converges in distribution to a standard normal distribution under the null under the conditions of Theorem 1. From the asymptotic normality of T_{LRST}, one can evaluate the statistical significance of the proposed test statistic T_{LRST}, based on the standard normal distribution. When the true $\bar{\theta}$ falls in the parameter space of the alternative hypothesis, we expect the proposed test statistic T_{LRST}, as the linear combination (with equal weights) of the rank differences across both outcomes and time points, would be large and the null hypothesis will be rejected.

Notice that the test of Huang et al.[2] is a special case of the longitudinal rank-sum test by only using the element $(\bar{R}_{y\cdot T\cdot} - \bar{R}_{x\cdot T\cdot})$ within $\boldsymbol{R}$. Let $J_T = (0, \cdots, 0, 1)$. Then, substituting J with J_T in $T_{LRST} = \frac{J^\top \boldsymbol{R}}{\sqrt{J^\top N \hat{\boldsymbol{\Sigma}} J}}$, the test statistic T_{LRST} reduces exactly to the test statistic of Huang et al.[2].

2.3.2 Main effect test when no interaction exists

When the interaction effect does not exist, namely, the relative treatment effect does not change across time, the analysis of multiple longitudinal outcome data is consequently simplified to analysis of two independent factors: main treatment effect and time effect. Since global treatment efficacy is of interest and no interaction exists, it is sufficient to test main treatment effect at any visit. The null and alternative of the main effect test in (2.6) reduces to:

$$H_0 : \theta_t = 0 \text{ for arbitrary } t, \quad vs \quad H_1 : \theta_t > 0 \tag{2.8}$$

Note that this hypothesis test is one-sided and it is recommended to set the type I error rate at 0.025, instead of 0.05, in a typical Phase III clinical trial, per FDA guideline.[30] Hypothesis (2.8) is the nonparametric Behrems-Fisher hypothesis problem under cross-sectional studies (at time t).

To test hypothesis (2.8), the test statistic would naturally be $\frac{\bar{R}_{y\cdot t\cdot} - \bar{R}_{x\cdot t\cdot}}{\sqrt{\widehat{var}(\bar{R}_{y\cdot t\cdot} - \bar{R}_{x\cdot t\cdot})}}$ for any $t \in \{1, \cdots, T\}$, where $\widehat{var}(\bar{R}_{y\cdot t\cdot} - \bar{R}_{x\cdot t\cdot})$ is basically $N\{\hat{\boldsymbol{\Sigma}}\}_{tt}$. In the absence of treatment and time interaction, $\frac{\bar{R}_{y\cdot t\cdot} - \bar{R}_{x\cdot t\cdot}}{\sqrt{\widehat{var}(\bar{R}_{y\cdot t\cdot} - \bar{R}_{x\cdot t\cdot})}}$ would give the same conclusion for arbitrary t when testing (2.8). Without loss of generality, we can use $T_{last} = \frac{\bar{R}_{y\cdot T\cdot} - \bar{R}_{x\cdot T\cdot}}{\sqrt{\widehat{var}(\bar{R}_{y\cdot T\cdot} - \bar{R}_{x\cdot T\cdot})}}$, which use the data from the last time point T, to test for the main treatment effect. T_{last} is exactly the test statistic proposed in Huang et al.[2], which only utilized the outcome data from the last observation. From the asymptotic

properties developed in Section 2.1 and the variance consistency of $\hat{\mathbf{\Sigma}}$, we have that T_{last} follows standard normal distribution asymptotically under the null hypothesis in (2.8). When the true θ_t falls in the parameter space of the alternative hypothesis, we expect the test statistic T_{last}, as the linear combination (with equal weights) of the rank differences across outcomes at time T, would be large and the null hypothesis will be rejected.

However, when no interaction exists, we suggest the proposed longitudinal rank-sum test statistic $T_{LRST} = \frac{\bar{R}_{y\cdots} - \bar{R}_{x\cdots}}{\sqrt{\widehat{var}(\bar{R}_{y\cdots} - \bar{R}_{x\cdots})}}$ as in Section 2.3.1 to test hypothesis in (2.8), it actually provides higher power than the Huang's test under alternative hypothesis. Please refer to Supplemental materials D for the detailed proof. Intuitively, the longitudinal rank-sum test has larger power because we employ the outcome information from a "larger" dataset (a dataset includes outcomes not only from the last visit, but also from other visits).

Therefore, to test the global treatment efficacy across the whole treatment period, the proposed longitudinal rank-sum test is an appropriate, yet powerful option regardless interaction exists or not, as compared to the existing cross-sectional global testing procedures.

3 Simulations

In this section, we conduct simulation studies to investigate the performance of the proposed rank-based test procedures for multivariate longitudinal outcomes in clinical trials. We first perform interaction test under various scenarios, and then perform both the Huang's test and the proposed longitudinal rank-sum test. Since the main treatment efficacy is of primary interest, we compare the proposed longitudinal rank-sum test to Huang's test in terms of type I error rate (under H_0) and power (under alternative) under various settings.

We assume there are three outcomes ($K = 3$) and four follow-up visits ($T = 4$, in addition to the baseline visit). We then generate the outcome values from a $15(K \times (T + 1))$-dimensional multivariate distribution. For both the control and treatment groups, we set the same population mean values for all the outcomes at baseline because of randomization. The population mean values

for the 3 outcomes across 5 time points (including the baseline) are considered as follows.

$$\boldsymbol{\mu}^{control} = \begin{pmatrix} .1 & .2 & .3 & .4 & .5 \\ .3 & .4 & .5 & .6 & .7 \\ .5 & .6 & .7 & .8 & .9 \end{pmatrix}, \boldsymbol{\mu}^1 = \begin{pmatrix} .1 & .4 & .5 & .6 & .7 \\ .3 & .6 & .7 & .8 & .9 \\ .5 & .8 & .9 & 1 & 1.1 \end{pmatrix}, \boldsymbol{\mu}^2 = \begin{pmatrix} .1 & .1 & .4 & .3 & .6 \\ .3 & .3 & .6 & .5 & .8 \\ .5 & .5 & .8 & .6 & 1.1 \end{pmatrix},$$

$$\boldsymbol{\mu}^3 = \begin{pmatrix} .1 & .3 & .5 & .7 & .8 \\ .3 & .5 & .8 & .8 & 1 \\ .5 & .7 & .9 & 1 & 1.2 \end{pmatrix}, \boldsymbol{\mu}^4 = \begin{pmatrix} .1 & .1 & .3 & .6 & .7 \\ .3 & .3 & .5 & .8 & .9 \\ .5 & .5 & .7 & 1 & 1 \end{pmatrix}, \boldsymbol{\mu}^5 = \begin{pmatrix} .1 & .4 & .5 & .5 & .5 \\ .3 & .6 & .7 & .7 & .7 \\ .5 & .8 & .9 & .9 & .9 \end{pmatrix},$$

where $\boldsymbol{\mu}^{control}$ is the population mean matrix of the control group, while $\boldsymbol{\mu}^1$ to $\boldsymbol{\mu}^5$ are the population mean matrices of the treatment group in five different scenarios. For each matrix, each row denotes one outcome's mean values across 5 time points and each column denotes multiple outcome values at a particular time point. For all scenarios, the mean values in $\boldsymbol{\mu}^1$ to $\boldsymbol{\mu}^5$ are arranged so that larger values over time imply the treatment efficacy in improving outcomes as compared to $\boldsymbol{\mu}^{control}$.

- Scenario 1 with population mean matrix $\boldsymbol{\mu}^1$ (interaction does not exist, main treatment effect exists): all outcomes improve over time.
- Scenario 2 with population mean matrix $\boldsymbol{\mu}^2$ (interaction exists, main treatment effect does not exist): all outcomes changes irregularly with the treatment having positive effect at the last time point.
- Scenario 3 with population mean matrix $\boldsymbol{\mu}^3$ (interaction exists, main treatment effect exists): all outcomes improve over time.
- Scenario 4 with population mean matrix $\boldsymbol{\mu}^4$ (interaction exists, main treatment effect exists): all outcomes deteriorate in the early stage, but the treatment has positive effect at the last time point.
- Scenario 5 with population mean matrix $\boldsymbol{\mu}^5$ (interaction exists, main treatment effect exist): all outcomes improve in the early stage, but there is no positive treatment effect at the last time point.

3.1 Multivariate Normal Data

We first generate data in the control group $\boldsymbol{x}_i = (x_{i01}, \cdots, x_{iT1}, \cdots, x_{i0K}, \cdots, x_{iTK})$, $i = 1, 2, \cdots, m$, i.i.d, from a 15-variate normal distribution with mean $vec(\boldsymbol{\mu}^{control})$ and covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_K \otimes \boldsymbol{\Sigma}_T$, where $vec(\boldsymbol{\mu})$ denotes vectorization of the matrix $\boldsymbol{\mu}$, $\otimes$ is the Kronecker product, an operation on two matrices of arbitrary sizes resulting in a block matrix, $\boldsymbol{\Sigma}_K$ is the covariance matrix of the outcome vector for each time point, with $\{\boldsymbol{\Sigma}_K\}_{kk} = 1$ for $k \in \{1, 2, 3\}$ and $\{\boldsymbol{\Sigma}_K\}_{ks} = 0.3$ for $k \neq s \in \{1, 2, 3\}$, $\boldsymbol{\Sigma}_T$ is the covariance matrix for a particular outcome across time, with $\{\boldsymbol{\Sigma}_T\}_{tt} = 1$ for $t \in \{1, 2, 3, 4, 5\}$ and $\{\boldsymbol{\Sigma}_T\}_{tr} = 0.6$ for $t \neq r \in \{1, 2, 3, 4, 5\}$. Similarly, under the null hypothesis, we generate data in the treatment group $\boldsymbol{y}_j = (y_{j01}, \cdots, y_{jT1}, \cdots, y_{j0K}, \cdots, y_{jTK})$, $j = 1, \cdots, n$, from the same distribution as for $\boldsymbol{x}_i$s. We expect that the probability of rejecting the null hypothesis for both interaction test (2.4) and main effect test (2.8) should be close to 0.05, suggesting that type I error is being controlled at the desired level. Under the alternative hypothesis, $\boldsymbol{y}_j$, $j = 1, \cdots, n$ is generated from the multivariate normal distribution with mean matrix being $\boldsymbol{\mu}^1$, $\boldsymbol{\mu}^2$, $\boldsymbol{\mu}^3$, $\boldsymbol{\mu}^4$, and $\boldsymbol{\mu}^5$, respectively, in Scenarios 1-5, and with the covariance matrix being $\boldsymbol{\Sigma}$.

We generate 10,000 replicates for each pair of m and n selected from $\{50, 100, 200\}$, rendering a total of nine combinations of sample sizes. For each replicate, we apply the proposed interaction test and longitudinal rank-sum test to each outcome's change from baseline to each time point, i.e., $\{\boldsymbol{x}_i^d\}_{K\times T}$ and $\{\boldsymbol{y}_i^d\}_{K\times T}$, where $\{\boldsymbol{x}_i^d\}_{kt} = x_{itk} - x_{i0k}$ and $\{\boldsymbol{y}_j^d\}_{kt} = y_{jtk} - y_{j0k}$, $t = 1, \cdots, T$, $k = 1, \cdots, K$. In comparison, Huang's test is applied to the changes from baseline to the last visit, which are the last column of $\{\boldsymbol{x}_i^d\}_{K\times T}$ and $\{\boldsymbol{y}_i^d\}_{K\times T}$.

The simulated type I error rate (under the null hypothesis) and the power (under the alternative hypothesis) are computed as the proportion of the null hypothesis H_0 being rejected at a nominal significance level of 0.05. The results are presented in Table 1. Under the null hypothesis of no interaction and no main treatment effect, the interaction test, Huang's test and the longitudinal rank-sum test have good control over the type I error (top portion of Table 1). Under the alternative hypothesis with mean matrices $\boldsymbol{\mu}^1$, the interaction test controls the type I error rate, and as expected, the longitudinal rank-sum test has larger power than Huang's test for the main treatment efficacy test. Under $\boldsymbol{\mu}^2$, the interaction test is significant, suggesting the treatment effect varies

over time and the test procedure based on a single time point is inappropriate. In this case, the longitudinal rank-sum test adequately controls the type I error, indicating no main treatment efficacy across the whole treatment period, while the Huang's test rejects the null hypothesis and erroneously declares treatment efficacy by only using the change from baseline to the last time point. Under $\boldsymbol{\mu}^3$ to $\boldsymbol{\mu}^5$, the interaction tests are all significant, suggesting the appropriateness of the proposed longitudinal rank-sum test. Under $\boldsymbol{\mu}^3$, both the Huang's test and longitudinal rank-sum test detect the treatment efficacy, with the longitudinal rank-sum test having slightly smaller power than Huang's test. This is reasonable because the longitudinal rank-sum test averages the treatment effect across all time points, while Huang's test only utilizes the changes from baseline to the last time point, which has artificially larger efficacy. Under the alternative hypothesis with $\boldsymbol{\mu}^4$, the Huang's test has much larger power than the longitudinal rank-sum test, because it "incorrectly" overlooks the deterioration of the outcomes in the early stage. Under the alternative hypothesis with $\boldsymbol{\mu}^5$, the longitudinal rank-sum test has markedly higher power than Huang's test, because it accounts for the outcome improvement before the end of study.

3.2 Log-normal Data

In this section, we examine the performance of the longitudinal rank-sum test under log-normal data. We first generate the multivariate normal data as in Section 3.1, transform the data into log-normal data by taking exponentiation, then compute each outcome's change from baseline to each time point. We run 10,000 simulations for each pair of m and n, where m and n are selected from $\{50, 100, 200\}$. The simulated type I error rate and the power calculations are displayed in Table 2. From Table 2, we have similar conclusions as for the multivariate normal data. Specifically, all three tests adequately control the type I error when there is no interaction and no main effect. The longitudinal rank-sum test has larger power than Huang's test under $\boldsymbol{\mu}^1$ when the interaction test is not significant. Under $\boldsymbol{\mu}^2$ to $\boldsymbol{\mu}^5$, all interaction tests are significant, suggesting the relative treatment effect is not parallel over time. In terms of main treatment effect, the longitudinal rank-sum test has good control of type I error under $\boldsymbol{\mu}^2$, while Huang's test has "inflated" power. Under $\boldsymbol{\mu}^3$, the longitudinal rank-sum test has slightly smaller power than Huang's test because it averages

Table 1: Type I error rate under the null hypothesis and power comparison results under the alternative hypothesis (under $\boldsymbol{\mu}^1$, $\boldsymbol{\mu}^2$, $\boldsymbol{\mu}^3$, $\boldsymbol{\mu}^4$, and $\boldsymbol{\mu}^5$) with a significance level of 0.05 (10,000 replicates) for the multivariate normal data.

(m,n)	(50,50)	(100,100)	(200,200)	(50,100)	(50,200)	(100,50)	(100,200)	(200,50)	(200,100)
Under the null (no interaction, no main treatment effect)									
All outcomes are the same for the two comparison groups									
Interaction test	0.059	0.053	0.049	0.057	0.066	0.059	0.051	0.059	0.054
Huang's test	0.055	0.051	0.050	0.050	0.052	0.050	0.053	0.057	0.050
Longitudinal rank-sum test	0.056	0.055	0.051	0.049	0.052	0.050	0.051	0.053	0.053
Under $\boldsymbol{\mu}^1$ (no interaction, main treatment effect exists)									
All outcomes improve over time for the treatment group									
Interaction test	0.061	0.054	0.050	0.057	0.066	0.060	0.052	0.060	0.053
Huang's test	0.465	0.688	0.915	0.525	0.613	0.530	0.796	0.611	0.797
Longitudinal rank-sum test	0.617	0.856	0.984	0.715	0.790	0.713	0.933	0.785	0.929
Under $\boldsymbol{\mu}^2$ (interaction exists, main treatment effect does not exist)									
All outcomes changes irregularly, but improve in the end									
Interaction test	0.554	0.855	0.993	0.677	0.761	0.675	0.944	0.776	0.941
Huang's test	0.273	0.419	0.639	0.307	0.363	0.313	0.499	0.367	0.501
Longitudinal rank-sum test	0.056	0.054	0.051	0.049	0.053	0.050	0.050	0.054	0.053
Under $\boldsymbol{\mu}^3$ (interaction exists, main treatment effect exists)									
All outcomes improve over time for the treatment group									
Interaction test	0.231	0.413	0.731	0.305	0.365	0.305	0.538	0.372	0.543
Huang's test	0.740	0.942	0.998	0.828	0.889	0.830	0.977	0.888	0.981
Longitudinal rank-sum test	0.672	0.904	0.993	0.774	0.845	0.773	0.957	0.837	0.959
Under $\boldsymbol{\mu}^4$ (interaction exists, main treatment effect exists)									
All outcomes worsen at the early stage of the treatment, but improve in the end									
Interaction test	0.570	0.879	0.995	0.708	0.798	0.718	0.959	0.801	0.957
Huang's test	0.364	0.556	0.805	0.410	0.492	0.413	0.658	0.488	0.661
Longitudinal rank-sum test	0.170	0.230	0.353	0.182	0.211	0.180	0.281	0.205	0.283
Under $\boldsymbol{\mu}^5$ (interaction exists, main treatment effect exists)									
All outcomes improve at the early stage of the treatment, but no effect in the end									
Interaction test	0.301	0.544	0.848	0.394	0.458	0.381	0.666	0.444	0.677
Huang's test	0.055	0.051	0.050	0.049	0.055	0.050	0.053	0.057	0.049
Longitudinal rank-sum test	0.349	0.526	0.773	0.392	0.463	0.396	0.626	0.452	0.629

Table 2: Type I error rate under the null hypothesis and power comparison results under the alternative hypothesis (under $\boldsymbol{\mu}^1$, $\boldsymbol{\mu}^2$, $\boldsymbol{\mu}^3$, $\boldsymbol{\mu}^4$, and $\boldsymbol{\mu}^5$) with a significance level of 0.05 (10,000 replicates) for the log-normal data.

(m,n)	(50,50)	(100,100)	(200,200)	(50,100)	(50,200)	(100,50)	(100,200)	(200,50)	(200,100)
Under the null (no interaction, no main treatment effect) All outcomes are the same for the two comparison groups									
Interaction test	0.058	0.057	0.054	0.063	0.066	0.063	0.052	0.060	0.054
Huang's test	0.057	0.049	0.050	0.055	0.052	0.052	0.054	0.055	0.051
Longitudinal rank-sum test	0.058	0.051	0.052	0.052	0.051	0.051	0.051	0.053	0.055
Under $\boldsymbol{\mu}^1$ (no interaction, main treatment effect exists) All outcomes improve over time for the treatment group									
Interaction test	0.059	0.058	0.054	0.062	0.066	0.062	0.056	0.061	0.056
Huang's test	0.486	0.722	0.935	0.574	0.645	0.568	0.835	0.640	0.830
Longitudinal rank-sum test	0.673	0.905	0.994	0.792	0.857	0.772	0.965	0.832	0.957
Under $\boldsymbol{\mu}^2$ (interaction exists, main treatment effect does not exist) All outcomes improve in the early stage, but deteriorate in the late stage									
Interaction test	0.481	0.788	0.981	0.610	0.690	0.607	0.895	0.706	0.901
Huang's test	0.295	0.445	0.677	0.337	0.393	0.339	0.532	0.380	0.526
Longitudinal rank-sum test	0.059	0.056	0.058	0.055	0.054	0.055	0.055	0.056	0.061
Under $\boldsymbol{\mu}^3$ (interaction exists, main treatment effect exists) All outcomes improve over time for the treatment group									
Interaction test	0.233	0.410	0.722	0.309	0.374	0.295	0.540	0.355	0.530
Huang's test	0.772	0.956	0.999	0.863	0.915	0.854	0.985	0.910	0.984
Longitudinal rank-sum test	0.738	0.947	0.998	0.849	0.905	0.831	0.983	0.888	0.978
Under $\boldsymbol{\mu}^4$ (interaction exists, main treatment effect exists) All outcomes worsen at the early stage of the treatment, but improve in the end									
Interaction test	0.507	0.811	0.984	0.639	0.729	0.640	0.919	0.731	0.921
Huang's test	0.387	0.584	0.836	0.445	0.521	0.441	0.700	0.512	0.688
Longitudinal rank-sum test	0.197	0.269	0.426	0.211	0.240	0.218	0.330	0.246	0.338
Under $\boldsymbol{\mu}^5$ (interaction exists, main treatment effect exists) All outcomes improve at the early stage of the treatment, but no effect in the end									
Interaction test	0.264	0.460	0.769	0.343	0.394	0.337	0.587	0.392	0.584
Huang's test	0.057	0.047	0.052	0.049	0.053	0.051	0.050	0.059	0.048
Longitudinal rank-sum test	0.375	0.564	0.816	0.437	0.507	0.431	0.687	0.497	0.677

the treatment effect across all time points, but it "correctly" has smaller power under $\boldsymbol{\mu}^4$ and much larger power under $\boldsymbol{\mu}^5$.

3.3 Ordinal Data

In this section, we evaluate the longitudinal rank-sum test on ordinal outcomes. We consider $T = 4$ and $K = 3$ ordinal outcomes with five different levels $(0, 1, 2, 3, 4)$ with higher level indicating clinically better values. To generate the ordinal data, we first generate the multivariate normal data as in Section 3.1, and then transform them into ordinal values based on the following rule:

$$x_{ikt} = I(c_1^k \leq x'_{ikt} < c_2^k) + 2I(c_2^k \leq x'_{ikt} < c_3^k) + 3I(c_3^k \leq x'_{ikt} < c_4^k) + 4I(x'_{ikt} > c_4^k), \tag{3.1}$$

where x_{ikt} is the ordinal observation for outcome k at time t for subject i in the control group, $k = 1, \cdots, 3$, $t = 0, 1, \cdots, 4$, and $\boldsymbol{c}^k = (c_1^k, c_2^k, c_3^k, c_4^k)^\top$ (with $c_1^k < c_2^k < c_3^k < c_4^k$) is the parameter vector that determines the distribution of ordinal outcome k. We set $\boldsymbol{c}^1 = (0.1, 0.4, 0.6, 0.9)^\top$, $\boldsymbol{c}^2 = (0.3, 0.6, 0.8, 1.1)^\top$, and $\boldsymbol{c}^3 = (0.5, 0.8, 1, 1.3)^\top$ so that each ordinal outcome has non-negligible portion in all levels. We then compute each outcome's change from baseline to each time point.

We run 10,000 simulations for each pair of m and n, where m and n are selected from $\{50, 100, 200\}$. The simulated type I error rate and the power comparisons are displayed in Table 3. We have similar conclusions as for the multivariate normal data. Specifically, all methods adequately control the type I error when the two comparison groups are exactly the same. The longitudinal rank-sum test obviously has larger power than Huang's test under $\boldsymbol{\mu}^1$ when there is no interaction. When the interaction exists under $\boldsymbol{\mu}^2$ to $\boldsymbol{\mu}^5$, Huang's test overestimates the treatment effect under $\boldsymbol{\mu}^2$ to $\boldsymbol{\mu}^4$ and underestimates the treatment effect under $\boldsymbol{\mu}^5$, while the longitudinal rank-sum test provides a reasonable estimate of the treatment effect.

In summary, when the outcomes are multivariate normal, log-normal, and ordinal, the interaction test and the longitudinal rank-sum test adequately controls the type I error. When no interaction exists, the longitudinal rank-sum test demonstrates better performance than Huang's test in detecting the treatment efficacy. When the interaction test is significant, the longitudinal rank-sum test, under various scenarios, provides reasonable assessment for the treatment efficacy while Huang's test gives misleading results because it does not account for the longitudinal trajectories.

4 A Real Data Example: The Azilect Study

We apply the proposed test procedures to the motivating Azilect (Rasagiline) study for evaluating the efficacy of rasagiline in multiple longitudinal outcomes. The Azilect study is a Phase 3, randomized, double-blind study to evaluate the efficacy and safety of rasagiline in Japanese patients with early PD.[6] Patients with a diagnosis of PD within 5 years were randomized 1:1 to receive rasagiline (1 mg/day) or placebo for up to 26 weeks. Please refer to Hattori et al.[6] for study details.

The Azilet study adopted as the primary endpoint the Movement Disorder Society Unified

Table 3: Type I error rate under the null hypothesis and power comparison results under the alternative hypothesis (under $\boldsymbol{\mu}^1$, $\boldsymbol{\mu}^2$, $\boldsymbol{\mu}^3$, $\boldsymbol{\mu}^4$, and $\boldsymbol{\mu}^5$) with a significance level of 0.05 (10,000 replicates) for the ordinal data.

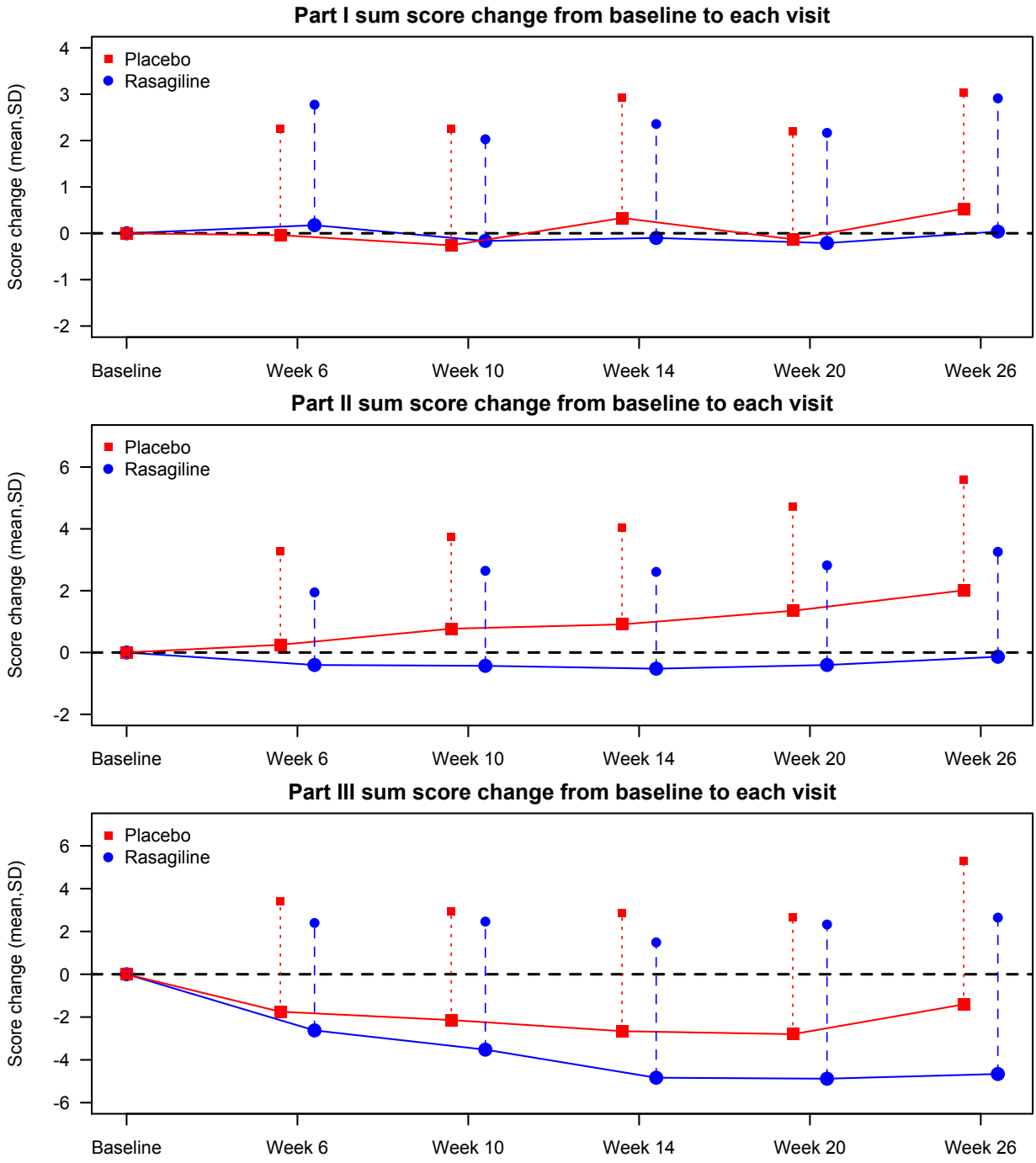
(m,n)	(50,50)	(100,100)	(200,200)	(50,100)	(50,200)	(100,50)	(100,200)	(200,50)	(200,100)
Under the null (no interaction, no main treatment effect)									
All outcomes are the same for the two comparison groups									
Interaction test	0.059	0.057	0.053	0.062	0.066	0.061	0.054	0.065	0.055
Huang's test	0.054	0.055	0.053	0.054	0.053	0.051	0.052	0.057	0.052
Longitudinal rank-sum test	0.056	0.049	0.052	0.051	0.053	0.050	0.051	0.054	0.053
Under $\boldsymbol{\mu}^1$ (no interaction, main treatment effect exists)									
All outcomes improve over time for the treatment group									
Interaction test	0.059	0.056	0.051	0.059	0.067	0.058	0.054	0.065	0.055
Huang's test	0.376	0.579	0.830	0.442	0.511	0.445	0.687	0.509	0.684
Longitudinal rank-sum test	0.529	0.771	0.958	0.631	0.699	0.618	0.872	0.691	0.863
Under $\boldsymbol{\mu}^2$ (interaction exists, main treatment effect does not exist)									
All outcomes improve in the early stage, but deteriorate in the late stage									
Interaction test	0.415	0.710	0.954	0.526	0.617	0.529	0.831	0.617	0.829
Huang's test	0.232	0.338	0.539	0.264	0.301	0.261	0.419	0.303	0.409
Longitudinal rank-sum test	0.055	0.051	0.052	0.051	0.055	0.051	0.053	0.055	0.054
Under $\boldsymbol{\mu}^3$ (interaction exists, main treatment effect exists)									
All outcomes improve over time for the treatment group									
Interaction test	0.179	0.315	0.579	0.231	0.287	0.230	0.414	0.271	0.408
Huang's test	0.625	0.867	0.985	0.717	0.796	0.722	0.939	0.790	0.935
Longitudinal rank-sum test	0.583	0.825	0.974	0.682	0.756	0.673	0.912	0.751	0.907
Under $\boldsymbol{\mu}^4$ (interaction exists, main treatment effect exists)									
All outcomes worsen at the early stage of the treatment, but improve in the end									
Interaction test	0.431	0.735	0.963	0.548	0.656	0.553	0.858	0.639	0.862
Huang's test	0.301	0.458	0.696	0.346	0.410	0.346	0.559	0.405	0.548
Longitudinal rank-sum test	0.154	0.202	0.315	0.170	0.189	0.167	0.245	0.187	0.245
Under $\boldsymbol{\mu}^5$ (interaction exists, main treatment effect exists)									
All outcomes improve at the early stage of the treatment, but no effect in the end									
Interaction test	0.228	0.401	0.700	0.293	0.340	0.292	0.512	0.348	0.521
Huang's test	0.053	0.054	0.054	0.049	0.055	0.049	0.051	0.056	0.055
Longitudinal rank-sum test	0.293	0.439	0.682	0.336	0.395	0.345	0.546	0.390	0.545

Parkinson's Disease Rating Scale (MDS-UPDRS), which is the most widely used scale for measuring parkinsonian symptoms in clinical and research practice.[31] MDS-UPDRS consists of 65 items, measured by a 5-point Likert scale (0-4, with higher values denoting an increased severity), in four parts: Part 1, Non-Motor Aspects of Experiences of Daily Living (13 items); Part 2, Motor Aspects of Experiences of Daily Living (13 items); Part 3, Motor Examination (33 items), and Part 4, Motor Complications (6 items). Please refer to Goetz et al.[32] for the details of the MDS-UPDRS scale. We apply the proposed tests to the multiple outcomes from the MDS-UPDRS scale in the following four settings: (1) MDS-UPDRS Part I sum score, Part II sum score, and Part III sum score (three outcomes, $K = 3$); (2) Part II sum score and Part III sum score (two outcomes, $K = 2$); (3) all 59 ordinal items from Parts I, II, and III ($K = 59$, with 13 Part I items, 13 Part II items, and 33 Part III items); and (4) all 46 ordinal items from Parts II and III ($K = 46$, with 13 Part II items and 33 Part III items).

A total 244 PD patients were randomized to receive placebo (n=126) or rasagiline (n=118). Patient characteristics at baseline were well balanced between two groups (Table 1 in Supplemental materials E). Moreover, there is no significant difference between the rasagiline and placebo groups at baseline in MDS-UPDRS Part I, Part II, and Part III sum scores.

In the placebo and rasagiline groups, there are a total of 100 and 109 patients who completed the baseline visit and all follow-up visits, respectively. We use the data from these 209 patients with complete MDS-UPDRS measurements as our analysis dataset. The MDS-UPDRS scale was measured at baseline, weeks 6, 10, 14, 20, 26. The mean changes of MDS-UPDRS sum score of each Part from baseline to each visit are displayed in Figure 1. The mean change of MDS-UPDRS Part I sum score from baseline to each visit is very close for both groups. Placebo group has steady increase (deterioration) in MDS-UPDRS Part II sum score, while the rasagiline group has some improvement during the study but the improvement seems to diminish at the end. Nevertheless, the treatment effect of rasagiline in MDS-UPDRS Part II sum score increases during the study, as compared to placebo. In MDS-UPDRS Part III sum score, both groups show steady decline (improvement) during the study while the rasagiline group has more improvement and increasing efficacy as compared to placebo.

Figure 1: The changes of MDS-UPDRS Part I, Part II, and Part III sum scores from baseline to each follow-up visit. Abbreviations: SD, standard deviation; MDS-UPDRS, The Movement Disorder Society Unified Parkinson's Disease Rating Scale.



Based on the analysis dataset of 209 patients, we apply the interaction test, longitudinal rank-sum test to the whole longitudinal data and apply Huang's test to the change from baseline to Week 26. The results of all four settings are displayed in Table 4. To ensure that higher scores reflect clinically better outcomes, as defined in Section 2, we multiply all ordinal item scores by -1 and obtain sum scores for Settings 1 and 2. In Setting 1 with three longitudinal outcomes (MDS-UPDRS Parts I, II, and III sum scores), we obtain $p = 0.021$ in interaction test, suggesting treatment effect changes over time, and the longitudinal rank-sum test should be used for detecting the main treatment efficacy. We get $p = 4.364 \times 10^{-6}$ in Huang's test and $p = 1.879 \times 10^{-4}$ in the longitudinal rank-sum test, suggesting statistical significance of rasagiline's global efficacy in these three outcomes. The p value from our proposed test is slightly larger than Huang's test. We believe that this is reasonable because our test evaluates the treatment efficacy by averaging the rank differences across all visits while Huang's test only utilizes the rank difference at the last visit. As suggested in Figure 1, while there may not be any treatment effect for Part I sum score, the treatment effect is comparatively smaller at the early stage but increasingly larger at the later stage for Parts II and III sum scores. In Setting 2 with two longitudinal outcomes (MDS-UPDRS Parts II and III sum scores), we obtain $p = 0.023$ in interaction test, which suggests the relative effect of rasagiline changes over time, $p = 1.485 \times 10^{-7}$ in Huang's test and $p = 1.211 \times 10^{-5}$ in the longitudinal rank-sum test, suggesting statistical significance of rasagiline's global efficacy in Parts II and III sum scores. As compared to Setting 1, the p values of both tests for main treatment effect in Setting 2 are smaller because of the significant treatment effect in Parts II and III, in comparison to no treatment effect in Part I (see below). Setting 2 is similar to the simulation scenario 3 with the population mean matrix $\boldsymbol{\mu}^3$, i.e., treatment improves all outcomes.

Alternatively, we apply all three tests to the longitudinal ordinal scores of MDS-UPDRS items. In Setting 3 with 59 ordinal items from Parts I, II, and III, we obtain $p = 0.002$ in interaction test, $p = 7.174 \times 10^{-7}$ in Huang's test and $p = 5.781 \times 10^{-5}$ in the longitudinal rank-sum test. In Setting 4 with 46 ordinal items from Parts II and III, we obtain $p = 0.003$ in interaction test, $p = 3.522 \times 10^{-7}$ in Huang's test and $p = 2.154 \times 10^{-5}$ in the longitudinal rank-sum test. Results from both settings suggest statistical significance of rasagiline's global efficacy in the longitudinal

Table 4: Results of the interaction test, longitudinal rank-sum test and Huang's test from various settings in the Azilect study. Abbreviations: MDS-UPDRS, The Movement Disorder Society Unified Parkinson's Disease Rating Scale.

Settings	Test	Test statistic	p value
MDS-UPDRS Part I, Part II and Part III sum scores ($K = 3$)	Interaction test	11.605	0.021
	Huang's test	-4.446	4.364×10^{-6}
	Longitudinal rank-sum test	-3.556	1.879×10^{-4}
MDS-UPDRS Part II and Part III sum scores ($K = 2$)	Interaction test	11.338	0.023
	Huang's test	-5.125	1.485×10^{-7}
	Longitudinal rank-sum test	-4.221	1.211×10^{-5}
59 ordinal outcomes from MDS-UPDRS Parts I, II, and III ($K = 59$)	Interaction test	17.370	0.002
	Huang's test	-4.820	7.174×10^{-7}
	Longitudinal rank-sum test	-3.855	5.781×10^{-5}
46 ordinal outcomes from MDS-UPDRS Parts II and III ($K = 46$)	Interaction test	16.111	0.003
	Huang's test	-4.960	3.522×10^{-7}
	Longitudinal rank-sum test	-4.090	2.154×10^{-5}

ordinal outcomes from the MDS-UPDRS items.

Finally, we apply all three tests to the longitudinal ordinal scores from each individual Part of MDS-UPDRS. The results are presented in Table 2 in Supplemental materials E. We use 13 ordinal scores in MDS-UPDRS Part I as the multivariate longitudinal outcomes, and obtain $p = 0.137$ in interaction test, $p = 0.078$ in Huang's test and $p = 0.276$ in the longitudinal rank-sum test, suggesting no significant interaction and main treatment effect in Part I. This conclusion is also supported by Figure 1 (upper panel), with the mean changes being very close in two groups. Analyzing the data from the 13 items in Part II and 33 items in Part III separately, we obtain $p = 0.041$ in interaction test, $p = 2.727 \times 10^{-6}$ in Huang's test and $p = 3.717 \times 10^{-5}$ in longitudinal rank-sum test in Part II, as compared to $p = 0.041$ in interaction test, $p = 2.459 \times 10^{-4}$ in Huang's test and $p = 2.673 \times 10^{-3}$ in longitudinal rank-sum test in Part III. The interaction tests in Part II and Part III suggests rasagiline efficacy varies over the treatment period. In terms of main treatments effect, these results indicate significant rasagiline efficacy in improving (reducing) patient perceived motor impact on function and clinician assessed severity of motor symptoms, captured by Part II and Part III MDS-UPDRS item scores, respectively. This conclusion is also substantiated by the visible group differences in Part II and Part III score changes displayed in Figure 1 (middle and lower panels).

5 Discussions

This paper develops a rank-based interaction test to detect the interaction between treatment effect and time effect, and then proposes a longitudinal rank-sum test to detect main treatment efficacy with multiple longitudinal outcomes. The test procedure objectively evaluates the effect of a new therapy longitudinally by fully utilizing the whole trajectories of multivariate longitudinal outcomes. In contrast, the traditional rank-based test procedures, such as the tests of O'Brien[1] and Huang et al.,[2] only utilize the information from the change from baseline to the last visit. Thus, it may render misleading conclusions as indicated in simulation scenarios under the population mean matrix $\boldsymbol{\mu}^2$ to $\boldsymbol{\mu}^5$. In the simulation studies, the interaction test and longitudinal rank-sum test have adequate control over the type I error and provide reasonable assessment for the treatment efficacy over time under various scenarios. The proposed test procedure is applied to the Azilect Study under various settings with different outcomes, and it demonstrates satisfactory performance by objectively detecting global treatment efficacy.

One major challenge in the longitudinal studies is the issue of missing data. Under missing at random (MAR) and missing not at random (MNAR) cases, ignoring the missing data may lead to biased estimates and invalid inference. To this end, multiple imputation and inverse probability weighting techniques can be used to handle missing data. By using appropriate multiple imputation techniques and models, multiple complete datasets are generated so that the interaction test and longitudinal rank sum test can be applied. Under weighting paradigm, the kernel function $\boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{y}_j)$ (defined in Supplemental materials A) of the U-statistic needs to be weighted by the estimated inverse probabilities to account for the missing outcome data. How to obtain consistent estimators of the response probabilities and incorporate them into the proposed longitudinal rank-sum test procedure is the subject of further research.

Also, based on the study design, we may adjust the weight of the multiple endpoints across time so that different weights can be assigned to different visits. This is clinically relevant as telemedicine has become standard practice in clinical care and research during the coronavirus disease pandemic. Research study participants may receive measurements of some health outcomes in both telemedicine and in-clinic visits. It may be necessary to assign larger weights to the in-clinic

visits than the telemedicine visits. To accommodate this, the test statistic T_{LRST} can be built based on $\boldsymbol{R}^{\boldsymbol{w}} = (w_1(\bar{R}_{y\cdot 1\cdot} - \bar{R}_{x\cdot 1\cdot}), \cdots, w_T(\bar{R}_{y\cdot T\cdot} - \bar{R}_{x\cdot T\cdot}))^\top$, where $w_t \geq 0$ and $\sum_{t=1}^{T} w_t = 1$, are chosen such that larger weight corresponds to more important clinical visits. When all weights are equal, this new test statistics reduces to the proposed longitudinal rank sum test.

Acknowledgement

The research of Sheng Luo was supported by National Institute on Aging (grant numbers: R01AG064803, P30AG072958, and P30AG028716).

The data that are used in our application study are available from the Critical Path Institute's CPP Consortium. The Critical Path Institute's CPP Consortium is funded by Parkinson's United Kingdom and the following industry members: AbbVie, Biogen, Cerevel, Denali, GSK, MSD, Takeda, Sanofi, Roche, IXICO, Cereval, Clario and UCB. We also acknowledge additional CPP member organizations, including the Parkinson's Disease Foundation, The Michael J. Fox Foundation, the Davis Phinney Foundation, The Cure Parkinson's Trust, PMD Alliance, the University of Oxford, University of Cambridge, Newcastle University, University of Glasgow, as well as the NINDS, US Food and Drug Administration, and the European Medicines Agency. We also acknowledge The Michael J. Fox Foundation for sponsoring of PPMI. Data were obtained from the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org/data). The PPMI is sponsored and partially funded by The Michael J. Fox Foundation for Parkinson's Research and funding partners, including AbbVie, Avid, Biogen, Bristol-Myers Squibb, Convance, GE Healthcare, Genentech, GSK, Lilly, Lundbeck, MSD, Meso Scale Discovery, Pfizer, Piramal, Roche, Sanofi Genzyme, Servier, TEVA, UCB, and Golub Capital. For up-to-date information on the study, visit www.ppmi-info.org. We would also like to recognize the scientific leadership of CPP advisors Karl Kieburtz, Tanya Simuni, Michael Schwarzschild, and Jesse Cedarbaum.

Data used in the preparation of this article were obtained from the CPP Unified Clinical Database. CPP acknowledges the contributions of UK investigators Michele Hu, Donald Grosset, Caroline Williams Gray, Rachael Lawson, and David Burn for their role in contributing data from PD cohort studies to the CPP Unified PD database.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

1 O'Brien PC. Procedures for Comparing Samples with Multiple Endpoints. Biometrics. 1984;40(4):1079-87.

2 Huang P, Tilley BC, Woolson RF, Lipsitz S. Adjusting O'Brien's Test to Control Type 1 Error for the Generalized Nonparametric Behrens-Fisher Problem. Biometrics. 2005;61(2):532-9.

3 Yang W, Hamilton JL, Kopil C, Beck JC, Tanner CM, Albin RL, et al. Current and projected future economic burden of Parkinson's disease in the U.S. Parkinson's Disease. 2020;6(15):1-9.

4 Parkinson Study Group STEADY-PD III Investigators. Isradipine Versus Placebo in Early Parkinson Disease: A Randomized Trial. Annals of Internal Medicine. 2020;172(9):591-8.

5 Parkinson Study Group SURE-PD3 Investigators. Effect of Urate-Elevating Inosine on Early Parkinson Disease Progression: The SURE-PD3 Randomized Clinical Trial. JAMA. 2021;326(10):926-39.

6 Hattori N, Takeda A, Takeda S, Nishimura A, Kitagawa T, Mochizuki H, et al. Rasagiline monotherapy in early Parkinson's disease: A phase 3, randomized study in Japan. Parkinsonism and Related Disorders. 2019;60:146-52.

7 O'Brien LM, Fitzmaurice GM. Analysis of Longitudinal Multiple-Source Binary Data Using Generalized Estimating Equations. Journal of the Royal Statistical Society Series C. 2004;53(1):177-93.

8 Bartholomew DJ. A test of homogeneity of means under order restrictions. Journal of the Royal Statistical Society, Series B. 1961;23(1):239-72.

9 Perlman MD. One-Sided Testing Problems in Multivariate Analysis. The Annals of Mathematical Statistics. 1969;40(2):549-67.

10 Robertson T, Wright FT, Dykstra RL. Order Restricted Statistical Inference. New York: John Wiley & Sons; 1988.

11 Tang DI. Uniformly More Powerful Tests in a One-Sided Multivariate Problem. Journal of the American Statistical Association. 1994;89:1006-11.

12 Silvapulle MJ, Sen P. Constrained Statistical Inference. Hoboken, New Jersey: Wiley; 2005.

13 Tang DI, Geller NL, Pocock SJ. On the Design and Analysis of Randomized Clinical Trials with Multiple Endpoints. Biometrics. 1993;49(1):23-30.

14 Bittman R, Romano JP, Vallarino C, Wolf M. Optimal testing of multiple hypotheses with common effect direction. Biometrika. 2009;96(2):399-410.

15 Lu ZH. Halfline tests for multivariate one-sided alternatives. Computational Statistics and Data Analysis. 2013;57:479-90.

16 Akritas MG, Brunner E. A unified approach to rank tests for mixed models. Journal of Statistical Planning and Inference. 1997;61:249 277.

17 Brunner E, Munzela U, Purib ML. The multivariate nonparametric Behrens–Fisher problem. Journal of Statistical Planning and Inference. 2002;108:37 53.

18 Brunner E, Konietschke F, Pauly M, Puri ML. Rank-based procedures in factorial designs: hypotheses about non-parametric treatment effects. Journal of the Royal Statistical Society Series B (Statistical Methodology). 2016;79(5):1463-85.

19 Roy A, Harrar S, Konietschke F. The nonparametric Behrens-Fisher problem with dependent replicates. Statistics in Medicine. 2019;38:4939-62.

20 Gunawardana A, Konietschke F. Nonparametric multiple contrast tests for general multivariate factorial designs. Journal of Multivariate Analysis. 2019;173:165-80.

21 Dobler D, Friedrich S, Pauly M. Nonparametric MANOVA in meaningful effects. Annals of the Institute of Statistical Mathematics. 2020;72:997-1022.

22 Liu A, Li Q, Liu C, Yu K, Yu KF. A Rank Test for Comparison of Multidimensional Outcomes. Journal of the American Statistical Association. 2010;105(490):578-87.

23 Zhang W, Liu A, Tang LL, Li Q. A cluster-adjusted rank-based test for a clinical trial concerning multiple endpoints with application to dietary intervention assessment. Biometrics. 2019;75(3):821-30.

24 Konietschke F, Bathke AC, Hothornc LA, Brunner E. Testing and estimation of purely nonparametric effects in repeated measures designs. Computational Statistics and Data Analysis. 2010;54:1895-905.

25 Zhuang Y, Guan Y, Qiu L, Lai M, Tan MT, Chen P. A novel rank-based non-parametric method for longitudinal ordinal data. Statistical Methods in Medical Research. 2018;27(9):2775-94.

26 Umlauft M, Placzek M, Konietschke F, Pauly M. Wild bootstrapping rank-based procedures: Multiple testing in nonparametric factorial repeated measures designs. Journal of Multivariate Analysis. 2019;171:176-92.

27 Rubarth K, Pauly M, Konietschke F. Ranking procedures for repeated measures designs with missing data: Estimation, testing and asymptotic theory. Statistical Methods in Medical Research. 2022;31(1):105-18.

28 Rubarth K, Sattler P, Zimmermann HG, Konietschke F. Estimation and Testing of Wilcoxon-Mann-Whitney Effects in Factorial Clustered Data Designs. Symmetry. 2022;14(244):1-34.

29 Brunner E, Domhof S, Langer F. Nonparametric Analysis of Longitudinal Data in Factorial Experiments. New York: Wiley; 2002.

30 U S Department of Health and Human Services, Food and Drug Administration, CDER&CBER. Multiple Endpoints in Clinical Trials: Guidance for Industry.

https://wwwfdagov/files/drugs/published/Multiple-Endpoints-in-Clinical-Trials-Guidance-for-Industrypdf. 2017.

31 Luo S, Goetz CG, Choi D, Aggarwal S, Mestre TA, Stebbins GT. Resolving Missing Data from the Movement Disorder Society Unified Parkinson's Disease Rating Scale: Implications for Telemedicine. Movement Disorder. 2022;37(8):1749-55.

32 Goetz CG, Tilley BC, Shaftman SR, Stebbins GT, Fahn S, Martinez-Martin P, et al. Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale Presentation and Clinimetric Testing Results. Movement Disorders. 2008;23(15):2129-70.