A novel longitudinal rank-sum test for multiple primary endpoints in clinical trials: Applications to neurodegenerative disorders

Xiaoming Xu, Dhrubajyoti Ghosh and Sheng Luo*

Department of Biostatistics and Bioinformatics
Duke University, Durham, NC, USA

May 8, 2024

Abstract

Neurodegenerative disorders such as Alzheimer's disease (AD) present a significant global health challenge, characterized by cognitive decline, functional impairment, and other debilitating effects. Current AD clinical trials often assess multiple longitudinal primary endpoints to comprehensively evaluate treatment efficacy. Traditional methods, however, may fail to capture global treatment effects, require larger sample sizes due to multiplicity adjustments, and may not fully exploit multivariate longitudinal data. To address these limitations, we introduce the Longitudinal Rank Sum Test (LRST), a novel nonparametric rank-based omnibus test statistic. The LRST enables a comprehensive assessment of treatment efficacy across multiple endpoints and time points without multiplicity adjustments, effectively controlling Type I error while enhancing statistical power. It offers flexibility against various data distributions encountered in AD research and maximizes the utilization of longitudinal data. Extensive simulations and real-data applications demonstrate the LRST's performance, underscoring its potential as a valuable tool in AD clinical trials. Nonparametrics, Global test, rank-sum-type test, U-Statistics.

1 Introduction

Alzheimer's disease (AD), a prevalent neurodegenerative disorder, affects millions globally, with a substantial 6.7 million cases in the United States alone (Association, 2023). It is characterized by progressive cognitive and functional deterioration, significantly impairing patient autonomy and caregiver quality of life. The etiology of AD involves a complex interplay of genetic, environmental, and lifestyle factors, leading to significant individual variability in its manifestation and progression (Scheiblich *and others*, 2020). This heterogeneity poses considerable challenges in developing efficacious treatments. In response, the US Food and Drug Administration (FDA) recommends incorporating both cognitive and functional outcomes as primary endpoints in AD clinical trials, in 2018 guidelines (Department of Health & Human Services *and others*, 2018). Current AD randomized clinical trials (RCTs) often capture multiple longitudinal primary endpoints, encompassing neuropsychological, functional, and behavioral assessments, and increasingly, blood-based biomarkers (Hansson *and others*, 2023). An analysis of 314 late-phase AD efficacy trials conducted since 1996 shows that nearly half (155 trials or 49.4%) included two or more primary endpoints, reflecting a shift towards a more comprehensive evaluation paradigm (Yu *and others*, 2022).

In evaluating multivariate longitudinal outcomes from AD RCTs that compare experimental treatments against placebos or active controls, two primary methodological approaches are prevalent. The first, the individual test approach, assesses longitudinal outcomes either independently or jointly to determine treatment effects for each outcome, followed by multiplicity adjustment. Analytical techniques include parametric methods such as univariate or multivariate GEE and linear mixed models (LMMs), as well as nonparametric methods such as univariate or multivariate functional mixed models (FMMs) (Zhu *and others*, 2019) and various rank-based tests (Akritas and Brunner, 1997; Konietschke *and others*, 2010; Brunner *and others*, 2017; Zhuang *and others*, 2018; Umlauft *and others*, 2019; Rubarth *and others*, 2022*a*,*b*). Multiplicity adjustment may employ co-primary endpoint strategy, requiring significance across all endpoints for treatment efficacy assertion, adopted in studies such as Semagacestat (Doody *and others*, 2013), EXPEDITION (Doody *and others*, 2014), LMTM (Gauthier *and others*, 2016), and

[0]Sheng Luo, email: sheng.luo@duke.edu

ADMET 2 (Mintzer *and others*, 2021), or multiple endpoint strategy, requiring at least one significant outcome after multiple comparison corrections (e.g., Bonferroni, Holm, or Hochberg procedures, and fixed-sequence methods) per FDA guidelines (GUIDANCE, 2022). However, these methods provide treatment effect estimates for individual outcomes, lacking a comprehensive assessment of global treatment efficacy across multiple endpoints when considering various disease aspects simultaneously. Additionally, Multiplicity adjustments can reduce power, requiring larger sample sizes (Hamasaki *and others*, 2018). Moreover, LMMs' treatment efficacy estimation may be biased by endpoint non-normality arising from skewness, outliers, and their combination (Lachos *and others*, 2011).

The second methodology in analyzing multivariate longitudinal outcomes in AD RCTs involves transforming longitudinal data into a cross-sectional format, focusing on changes from baseline to the final observation. This approach typically employs global testing procedures, which are broadly categorized into parametric and nonparametric methods. Parametric procedures, reliant on multivariate normality, include likelihood ratio tests (Bartholomew, 1961; Perlman, 1969; Tang, 1994) and t-statistics type tests (Tang *and others*, 1993; Bittman *and others*, 2009; Lu, 2013). However, their practical use is often limited due to stringent distributional assumptions. In contrast, nonparametric rank-based procedures, which are distribution-free, are widely used in clinical research settings, including multivariate factorial designs (Munzel and Brunner, 2000; Gunawardana and Konietschke, 2019), multivariate categorical data (Brunner *and others*, 2002*b*), cluster-correlated data (Larocque *and others*, 2010), subjects with dependent replicates (Roy *and others*, 2019), multivariate crossed layouts (Dobler *and others*, 2020), and high-dimensional genetic data (Meng *and others*, 2022). Specifically, Brunner *and others* (2002*b*) developed Wald-type and ANOVA-type tests for simultaneous multiple outcome assessments. Roy *and others* (2019) addressed nonparametric methods for two-sample problems with correlated replicates. O'Brien (1984) proposed a rank-sum-type test, which was later refined by Huang *and others* (2005) through a variance adjustment to control Type I error rate. Liu *and others* (2010) introduced a rank-max-type test, effective in detecting group differences across all outcomes, particularly when these differences are directionally opposite. Additionally, Zhang *and others* (2019) developed a cluster-adjusted rank-based test, combining the strengths of both rank-sum and rank-max tests by aggregating within-cluster evidence and combining it at the cluster level. Despite these advancements, these rank-based tests are primarily applied to cross-sectional data and cannot fully utilize multivariate longitudinal data. Consequently, they are inadequate for tracking temporal variations in treatment efficacy, which is crucial for a comprehensive assessment of treatment effects over the entire treatment period.

Addressing these gaps, this article introduces the Longitudinal Rank Sum Test (LRST), a novel nonparametric rank-based omnibus test designed to comprehensively evaluate treatment efficacy across multiple longitudinal primary endpoints in AD RCTs. The LRST offers a robust and efficient statistical framework for detecting treatment effects across multiple endpoints over time. Compared to existing methods, the LRST innovates in several key areas. Firstly, it enables a global assessment of treatment efficacy across multiple endpoints and time points without requiring multiplicity adjustments. This approach enhances understanding of treatment effects and provides a holistic perspective on interventions in AD trials. Secondly, the LRST effectively controls Type I error while enhancing statistical power, thereby requiring smaller sample sizes. This efficiency reduces subjects' exposure to potentially ineffective treatments and lowers the financial burden of therapy development in AD RCTs. Thirdly, the LRST employs a nonparametric approach based on ranking, akin to the Wilcoxon rank-sum test, without relying on distributional assumptions such as normality in LMMs. This flexibility ensures robustness against various data distributions commonly encountered in AD research. Lastly, the LRST procedure maximizes the utilization of all longitudinal data, capturing dynamic changes across multiple endpoints and providing a comprehensive analysis of treatment effects in AD RCTs.

The rest of the article is organized as follows. Section 2 develops LRST procedure and examines its theoretical properties. In Section 3, extensive simulations are conducted to investigate the performance in Type I error and power. In Section 4, we applied the LRST procedure to two RCTs of neurodegenerative disorder: the Bapineuzumab (Bapi) 301 AD study and the Azilect Parkinson's disease (PD) study. Some discussions and future work are given in Section 5. All technical details and additional results are provided in the Supplementary material.

2 Statistical Framework

2.1 Notation and Asymptotic Distribution

In AD RCTs with multiple longitudinal outcomes and two parallel arms, treatment vs. control, the data structure is often of the type (x_{itk}, y_{jtk}, t), where x_{itk} is the change of outcome k ($k = 1, \ldots K$ for a total of K outcomes) from baseline to time t ($t = 1, \ldots T$, where T is the number of after-baseline visits) from subject i ($i = 1, \ldots n_x$) in the control group, while y_{jtk} is defined similarly for subject j ($j = 1, \ldots n_y$) in the treatment group, n_x and n_y

are the sample sizes of the control and treatment groups, respectively. For simplicity, we let larger outcome values be clinically favorable. We combined two vectors $(x_{1tk}, \dots x_{n_x tk})$ and $(y_{1tk}, \dots y_{n_y tk})$, which are the control and treatment groups' observations from outcome k at time t, rank them, with larger values obtaining higher ranks, and obtain R_{xitk} and R_{yjtk} being the mid-ranks of observations x_{itk} and y_{jtk}, respectively. The midrank is either the regular rank when there is no tie on the observations or the average rank of those tied observations. Further, we define:

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

and

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

for X-sample and Y-sample, respectively.

To build the LRST procedure, we followed Brunner *and others* (2002*a*) and define $\theta_{tk} = P(X_{tk} < Y_{tk}) - P(X_{tk} > Y_{tk})$ as the relative treatment effect in comparison to the control on outcome k at time t, where X_{tk} and Y_{tk} are the random variables corresponding to their observed values x_{tk} and y_{tk}. We let $\theta_t = \frac{1}{K}\sum_{k=1}^{K}\theta_{tk}$ be the relative treatment effect across all outcomes at time t. Notice that $\hat{\theta}_{tk} = \frac{1}{n_x n_y}\sum_{i=1}^{n_x}\sum_{j=1}^{n_y}[I(x_{itk} < y_{jtk}) - I(x_{itk} > y_{jtk})] = \frac{2}{N}(\bar{R}_{y\cdot tk} - \bar{R}_{x\cdot tk})$ is a consistent estimator of θ_{tk} (see the detailed derivations in Section S1 of the Supplementary material). 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 θ_t.

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

$$\begin{aligned}\frac{1}{\sqrt{N}}\begin{pmatrix}\bar{R}_{y\cdot 1\cdot} - \bar{R}_{x\cdot 1\cdot}\\ \bar{R}_{y\cdot 2\cdot} - \bar{R}_{x\cdot 2\cdot}\\ \vdots\\ \bar{R}_{y\cdot T\cdot} - \bar{R}_{x\cdot T\cdot}\end{pmatrix} &= \frac{\sqrt{N}}{2n_x n_y}\sum_{i=1}^{n_x}\sum_{j=1}^{n_y}\begin{pmatrix}\frac{1}{K}\sum_{k=1}^{K}[I(x_{i1k} < y_{j1k}) - I(x_{i1k} > y_{j1k})]\\ \frac{1}{K}\sum_{k=1}^{K}[I(x_{i2k} < y_{j2k}) - I(x_{i2k} > y_{j2k})]\\ \vdots\\ \frac{1}{K}\sum_{k=1}^{K}[I(x_{iTk} < y_{jTk}) - I(x_{iTk} > y_{jTk})]\end{pmatrix}\\ &= \frac{\sqrt{N}}{2n_x n_y}\sum_{i=1}^{n_x}\sum_{j=1}^{n_y}\boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{y}_j),\end{aligned}$$

where $\boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{y}_j)$ is the kernel function of the above multivariate 2 sample U-statistics and $\boldsymbol{x}_i = (x_{i11}, \cdots, x_{iTK})^\top$, $\boldsymbol{y}_j = (y_{j11}, \cdots, y_{jTK})^\top$, $\boldsymbol{h}_t(\boldsymbol{x}_i, \boldsymbol{y}_j) = \frac{1}{K}\sum_{k=1}^{K}[I(x_{itk} < y_{jtk}) - I(x_{itk} > y_{jtk})]$. By the asymptotic properties of the U-statistics, $\frac{1}{\sqrt{N}}(\bar{R}_{y\cdot 1\cdot} - \bar{R}_{x\cdot 1\cdot}, \cdots, \bar{R}_{y\cdot T\cdot} - \bar{R}_{x\cdot T\cdot})^\top$ has multivariate normal distribution asymptotically. We established 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 $\min(n_x, n_y) \to \infty$ and $\frac{n_x}{n_y} \to \lambda < \infty$, where $\{\boldsymbol{\Sigma}\}_{t_1 t_2}$ can be approximated by:*

$$\{\boldsymbol{\Sigma}\}_{t_1 t_2} \approx \frac{1}{K^2}\sum_{k_1=1}^{K}\sum_{k_2=1}^{K}\left[\left(1 + \frac{1}{\lambda}\right)c_{t_1 k_1, t_2 k_2} + (1+\lambda)\, d_{t_1 k_1, t_2 k_2}\right], \tag{1}$$

where $c_{t_1 k_1, t_2 k_2} = cov(G_{t_1 k_1}(X_{t_1 k_1}), G_{t_2 k_2}(X_{t_2 k_2}))$ and $d_{t_1 k_1, t_2 k_2} = cov(F_{t_1 k_1}(Y_{t_1 k_1}), F_{t_2 k_2}(Y_{t_2 k_2}))$.

Please refer to Section S1 of the Supplementary material for detailed proof. From the expression of $\{\boldsymbol{\Sigma}\}_{t_1 t_2}$ in Equation (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 (1),

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

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

and

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

account for the intra-source correlation (same outcome at different visits), inter-source correlation (different outcomes at the same visit), and cross-correlation (different outcomes at different visits), respectively. Theorem 2 gives the moment estimates of $c_{t_1 k_1, t_2 k_2}$ and $d_{t_1 k_1, t_2 k_2}$, and the proof is given in Section S2 of the Supplementary material.

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*

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

and

$$\hat{d}_{t_1k_1,t_2k_2} = \frac{1}{n_y}\sum_{j=1}^{n_y}\left\{\left[\frac{1}{n_x}\sum_{i=1}^{n_x} I(x_{it_1k_1} < y_{jt_1k_1}) - \frac{1+\hat{\theta}_{t_1k_1}}{2}\right]\left[\frac{1}{n_x}\sum_{i=1}^{n_x} I(x_{it_2k_2} < y_{jt_2k_2}) - \frac{1+\hat{\theta}_{t_2k_2}}{2}\right]\right\}, \quad (3)$$

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

2.2 Computation of $\hat{\boldsymbol{\Sigma}}$

We now calculate $\hat{\boldsymbol{\Sigma}}$, the estimate of the covariance matrix $\boldsymbol{\Sigma}$. For $t_1, t_2 \in \{1, 2, \cdots, T\}$, from Theorem 1, the t_1t_2th element of $\boldsymbol{\Sigma}$ is given by:

$$\begin{aligned}\{\boldsymbol{\Sigma}\}_{t_1t_2} &= cov\left(\frac{1}{\sqrt{N}}(\bar{R}_{y\cdot t_1\cdot} - \bar{R}_{x\cdot t_1\cdot}), \frac{1}{\sqrt{N}}(\bar{R}_{y\cdot t_2\cdot} - \bar{R}_{x\cdot t_2\cdot})\right)\\ &\approx \frac{1}{K^2}\sum_{k_1=1}^{K}\sum_{k_2=1}^{K}\left(\left(1+\frac{1}{\lambda}\right)c_{t_1k_1,t_2k_2} + (1+\lambda)\,d_{t_1k_1,t_2k_2}\right).\end{aligned}$$

Plugging in the moment estimates $\hat{c}_{t_1k_1,t_2k_2}$ and $\hat{d}_{t_1k_1,t_2k_2}$ in Theorem 2 for $k_1, k_2 \in \{1, 2, \cdots, K\}$, we obtain the covariance estimate $\hat{\boldsymbol{\Sigma}}$.

For ease of computation, let $R_y(x_{itk})$ be the mid-rank of x_{itk} among $\{y_{1tk}, y_{2tk}, \cdots, y_{n_ytk}, x_{itk}\}$ and $R_x(y_{jtk})$ be the mid-rank of y_{jtk} among $\{x_{1tk}, x_{2tk}, \cdots, x_{n_xtk}, y_{jtk}\}$. Also, we define:

$$\hat{\theta}_{tk} = \frac{1}{n_xn_y}\sum_{i=1}^{n_x}\sum_{j=1}^{n_y}[I(x_{itk} < y_{jtk}) - I(x_{itk} > y_{jtk})],$$

for $t = 1, 2, \cdots, T$, $k = 1, 2, \cdots, K$. Then, $\hat{\theta}_{tk}$ is a consistent estimator of θ_{tk} by the law of large numbers.

Let $\boldsymbol{P}_t = \left(P_{ik}^{(t)}\right)_{n_x\times K}$ with $P_{ik}^{(t)} = R_y(x_{itk}) - 1 - n_y(1-\hat{\theta}_{tk})/2$ and $\boldsymbol{Q}_t = \left(Q_{jk}^{(t)}\right)_{n_y\times K}$ with $Q_{jk}^{(t)} = R_x(y_{jtk}) - 1 - n_x(1+\hat{\theta}_{tk})/2$, $i = 1, 2, \cdots, n_x$, $j = 1, 2, \cdots, n_y$. Then, an empirical estimate of $c_{t_1k_1,t_2k_2}$ is

$$\hat{c}_{t_1k_1,t_2k_2} = \frac{1}{n_xn_y^2}\sum_{i=1}^{n_x} P_{ik_1}^{(t_1)}P_{ik_2}^{(t_2)},$$

which is the expression in Equation (2). Similarly, the estimate of $d_{t_1k_1,t_2k_2}$ is given by:

$$\hat{d}_{t_1k_1,t_2k_2} = \frac{1}{n_x^2n_y}\sum_{j=1}^{n_y} Q_{jk_1}^{(t_1)}Q_{jk_2}^{(t_2)},$$

which has the expression in Equation (3). Further, define $\boldsymbol{C}_{t_1t_2} = (c_{t_1k_1,t_2k_2})_{K\times K}$ and $\boldsymbol{D}_{t_1t_2} = (d_{t_1k_1,t_2k_2})_{K\times K}$, $t_1, t_2 = 1, 2, \cdots, T$, $k_1, k_2 = 1, 2, \cdots, K$. Then, it follows that the consistent estimators of $\boldsymbol{C}_{t_1t_2}$ and $\boldsymbol{D}_{t_1t_2}$ are $\hat{\boldsymbol{C}}_{t_1t_2} = \boldsymbol{P}_{t_1}^\top\boldsymbol{P}_{t_2}/(n_xn_y^2)$ and $\hat{\boldsymbol{D}}_{t_1t_2} = \boldsymbol{Q}_{t_1}^\top\boldsymbol{Q}_{t_2}/(n_x^2n_y)$, with their elements being $\hat{c}_{t_1k_1,t_2k_2}$ and $\hat{d}_{t_1k_1,t_2k_2}$, respectively. Then, $\{\hat{\boldsymbol{\Sigma}}\}_{t_1t_2}$ can be written as:

$$\{\hat{\boldsymbol{\Sigma}}\}_{t_1t_2} = \frac{1}{K^2}\left[\left(1+\frac{1}{\lambda}\right)\cdot sum(\hat{\boldsymbol{C}}_{t_1t_2}) + (1+\lambda)\cdot sum(\hat{\boldsymbol{D}}_{t_1t_2})\right],$$

where $sum(\hat{\boldsymbol{C}}_{t_1t_2})$ and $sum(\hat{\boldsymbol{D}}_{t_1t_2})$ denote the summation of all the elements of $\hat{\boldsymbol{C}}_{t_1t_2}$ and $\hat{\boldsymbol{D}}_{t_1t_2}$, respectively.

2.3 The longitudinal rank sum test

We let $\bar{\theta} = \frac{1}{T}\sum_{t=1}^{T}\theta_t$ be the overall treatment efficacy across all outcomes and time points. If the treatment improves on or slows the progression of some outcomes, we would expect to have $\bar{\theta} > 0$. Hence, we tested the overall treatment efficacy with the following null and alternative hypotheses.

$$H_0 : \bar{\theta} = 0 \quad vs \quad H_1 : \bar{\theta} > 0. \tag{4}$$

For testing the hypothesis in (4), we proposed the longitudinal rank-sum test (LRST) statistic

$$T_{LRST} = \frac{\bar{R}_{y\cdots} - \bar{R}_{x\cdots}}{\sqrt{\widehat{var}(\bar{R}_{y\cdots} - \bar{R}_{x\cdots})}}, \tag{5}$$

where $\bar{R}_{y\cdots} = \frac{1}{n_y TK}\sum_{j=1}^{n_y}\sum_{k=1}^{K}\sum_{t=1}^{T} R_{yjtk}$, $\bar{R}_{x\cdots} = \frac{1}{n_x TK}\sum_{i=1}^{n_x}\sum_{k=1}^{K}\sum_{t=1}^{T} R_{xitk}$ are the average rank across all outcomes and time points in the treatment and control groups, respectively. Larger values of the group rank difference $\bar{R}_{y\cdots} - \bar{R}_{x\cdots}$ suggest treatment efficacy as compared with the control. The denominator, $\widehat{var}(\bar{R}_{y\cdots} - \bar{R}_{x\cdots})$ is a consistent estimator of the variance of group rank difference $\bar{R}_{y\cdots} - \bar{R}_{x\cdots}$. The LRST statistic is an advanced rank-sum-type test that enhances existing rank-based tests (O'Brien, 1984; Huang *and others*, 2005; Liu *and others*, 2010; Zhang *and others*, 2019) by effectively leveraging all available longitudinal data. The non-rejection of the null hypothesis H_0, as defined in Equation (4), using the LRST statistic outlined in Equation (5), implies the absence of an overall treatment effect across multiple longitudinal outcomes. Conversely, the rejection of H_0 suggests that the treatment demonstrates greater effectiveness compared to the control group, providing directional inference. Upon rejection, one may utilize LMMs for each endpoint to identify and quantify treatment effects on specific outcomes.

Next, we derived the asymptotic distribution of the LRST statistic. 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. Incorporating the consistent covariance estimator $\hat{\boldsymbol{\Sigma}}$, we established 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 hypothesis, following the conditions specified in 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. Notice that the test of Huang *and others* (2005) 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 $\boldsymbol{R}$ because it transforms longitudinal data into a cross-sectional format, focusing on changes from baseline to the final observation. 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 *and others* (2005).

In the simulation and real data analysis, We assessed the performance of the LRST with two other inference methods: (1) the nonparametric factorial ANOVA rank statistics (NFARS) (Brunner *and others*, 2017), implemented using `R` package `nparLD` (Noguchi *and others*, 2012), which conducts multiple two-sample univariate tests and generates p-values for each outcome; (2) the univariate LMM, which incorporates covariates including time, treatment, their interaction, with the overall treatment effect determined via the F-test. For both alternative methods, we applied Bonferroni's correction for multiple comparisons, declaring significance when at least one univariate test's p-value fell below the Bonferroni-adjusted significance threshold.

3 Simulation study

3.1 Simulation setting

Our comprehensive simulation study was designed to rigorously evaluate the LRST test's capability in controlling Type I error rates and its power to detect significant overall treatment effects. The study parameters and structure were modeled after the Bapi 302 clinical trial (as described in Section 4.1), incorporating identical primary endpoints (ADAS-cog11 and DAD scores), trial progression patterns, duration, assessment schedules, and a placebo to treatment allocation ratio of 2:3. To simulate the placebo arm data, we first derived from the Bapi 302 study the mean changes from baseline to subsequent visits in ADAS-cog11 and DAD scores, along with their standard deviations (SD), as displayed in Table 1. For each time interval, data were generated using a piecewise linear mixed model with the mean changes and SDs in Table 1. We modeled within-subject temporal variability by integrating random effects with standard deviations matching those observed in the study, employing an AR(1) structure with a lag-1 autocorrelation coefficient of 0.6.

Visit Week	13	26	39	52	65	78
ADAS-cog11	$0.739_{4.799}$	$1.322_{5.386}$	$3.166_{6.510}$	$4.607_{7.444}$	$5.899_{8.084}$	$7.457_{9.139}$
DAD	$-0.706_{10.561}$	$-4.065_{13.057}$	$-5.705_{14.960}$	$-8.249_{15.662}$	$-12.104_{16.940}$	$-13.941_{18.080}$

Table 1: Mean changes from baseline to each visit for ADAS-cog11 and DAD scores, with standard deviations in subscripts. Abbreviation: ADAS-cog11, Alzheimer's Disease Assessment Scale - cognitive subscale (11 items); DAD, Disability Assessment for Dementia.

Furthermore, recognizing the LRST's capacity to analyze both continuous and ordinal outcomes, our study extends its application to ordinal data, which often presents extensive ties. To investigate the test's robustness under such conditions, we transformed simulated continuous data into ordinal categories (0 through 4). Specifically, continuous simulated data were discretized into ordinal categories (0, 1, 2, 3, 4), utilizing thresholds set at $\mu - 3\sigma, \mu - \sigma, \mu + \sigma, \mu + 3\sigma$, where μ and σ reference the means and standard deviations from Table 1. The application of LRST to ordinal data derived from continuous simulations is denoted as LRST-ordinal. This extension allows us to assess the LRST's adaptability and performance in handling the complexities associated with multiple ordinal longitudinal outcomes in clinical research settings.

3.2 Simulation results

3.2.1 Type I error

To assess Type I error, we conducted simulations assuming no treatment effects across both outcomes and throughout all visits for both groups. To introduce inter-source correlation between the two outcomes, we set the between-outcome correlation coefficient as 0.5. The empirical Type I error rates of the LRST, at nominal levels of $\alpha = 0.05$ and 0.1, across various sample sizes (N) from 1000 simulations, are presented in Table 2. These results closely align with the specified α levels, demonstrating the LRST's capacity to maintain accurate Type I error rates, even in scenarios with smaller sample sizes, e.g., $N = 100$. Moreover, when the continuous simulated data were converted into ordinal outcomes, the LRST effectively controlled Type I error, underscoring its adeptness at managing the complexities introduced by significant ties in multivariate longitudinal ordinal data.

	LRST		LRST-ordinal	
N	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.1$
100	0.058	0.104	0.055	0.096
300	0.047	0.099	0.050	0.106
500	0.049	0.100	0.054	0.109
700	0.053	0.093	0.054	0.107
900	0.047	0.092	0.051	0.099
1200	0.059	0.106	0.054	0.097
1500	0.530	0.101	0.051	0.094

Table 2: Comparison of empirical Type I error rates between LRST for continuous data and LRST-ordinal for discretized data across various sample sizes (N) and nominal levels of α. The placebo-to-treatment allocation ratio was 2:3. The results are based on 1,000 simulations.

3.2.2 Power

In our power analysis, we considered two scenarios. In Scenario 1, we assumed that the Bapi group would exhibit the pre-specified effect sizes (2.21 unit advantage in ADAS-cog11 and 5.38 unit advantage in DAD over placebo, equally distributed across each visit) and set the between-outcome correlation coefficient as 0.5. We investigated the power under various total sample sizes ($N = 100, 300, 500, 700, 900, 1200, 1500$) with a fixed allocation ratio of 2:3 to the placebo and treatment groups. In Scenario 2, we fixed the total sample size at $N = 759$ (mirroring the actual sample sizes of the placebo ($n_x = 311$) and Bapi ($n_y = 448$) groups in our final analysis dataset). We varied the effect size to range from 0.2 to 2 times that specified in Scenario 1 (denoted as "0.2 to 2 deviations") and set the between-outcome correlation coefficient to be 0, 0.2, 0.5, and 0.8. In both scenarios, we conducted 1,000 simulations of the RCT.

Figure 1 displays the empirical power curves of the LRST, NFARS, and LMM methods across different total sample sizes in Scenario 1. The proposed LRST test consistently has higher power than the NFARS and LMM

tests across all sample sizes considered. For example, at $N = 900$, LRST reached a power of 0.863, significantly surpassing the LMM and NFARS with powers of 0.584 and 0.548, respectively. This result highlights the efficacy of the LRST in evaluating overall treatment impact across multiple longitudinal outcomes, proving to be more powerful than individual univariate parametric or nonparametric tests that require adjustments for multiple comparisons. Notably, the LMM method exhibited marginally greater power compared to NFARS across most sample sizes, due to the simulation data being generated under the piecewise LMM framework. Table 3 provides a comparison of empirical power between LRST applied to continuous data and LRST on ordinal data discretized from the same continuous simulations in Scenario 1. Despite the introduction of extensive ties through discretization, the LRST applied to ordinal data exhibits slightly lower yet remarkably close power to that of its continuous counterpart, confirming the LRST's robust performance in scenarios characterized by significant ties in multivariate longitudinal ordinal outcomes.

Figure 2 presents the empirical power curves of the LRST, NFARS, and LMM methods under various effect sizes and inter-outcome correlations in Scenario 2. Throughout these configurations, the LRST consistently demonstrates significantly greater power than both LMM and NFARS, underscoring its superior efficacy in detecting treatment effects across various sample sizes and correlations. This confirms LRST's advantage in evaluating treatment effects across multiple longitudinal outcomes, surpassing the capabilities of univariate tests adjusted for multiple comparisons. LMM's marginally higher power relative to NFARS is again due to the simulation framework favoring the LMM approach.

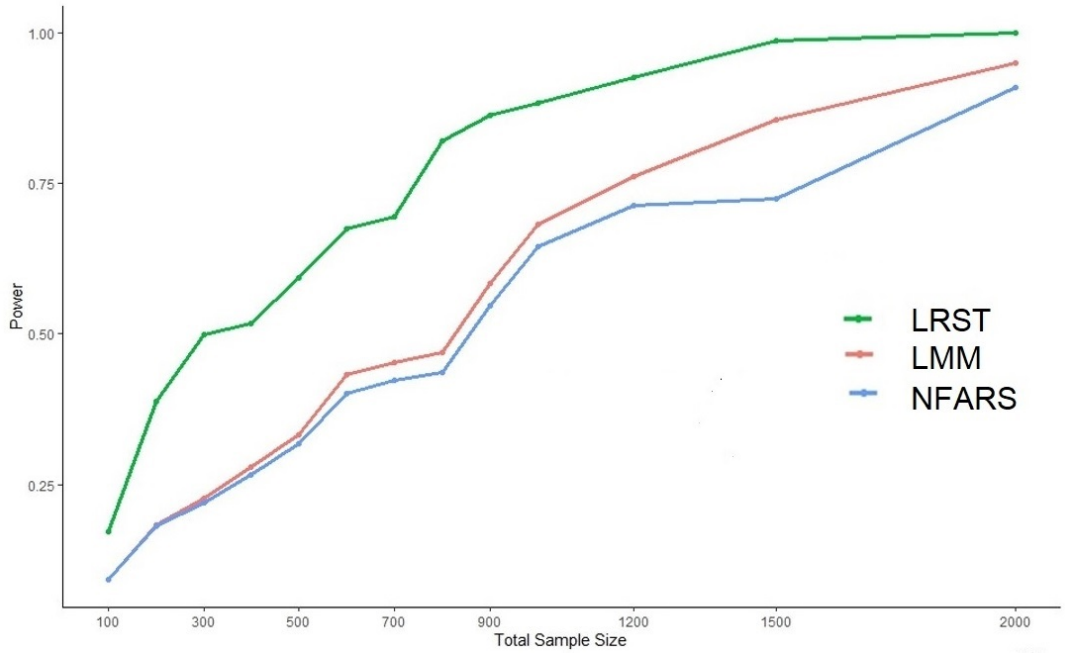

Figure 1: Empirical power curves of LRST, NFARS, and LMM methods for varying total sample sizes N in Scenario 1. The placebo-to-treatment allocation ratio was 2:3. Effect sizes were preset at 2.21 for ADAS-cog11 and 5.38 for DAD. Abbreviation: NFARS, the nonparametric factorial ANOVA rank statistics (Brunner *and others*, 2017); LMM, linear mixed model; ADAS-cog11, Alzheimer's Disease Assessment Scale - cognitive subscale (11 items); DAD, Disability Assessment for Dementia.

N	100	300	500	700	900	1200	1500
LRST	0.171	0.499	0.593	0.726	0.863	0.926	0.987
LRST-ordinal	0.172	0.481	0.584	0.696	0.823	0.893	0.929

Table 3: Comparison of empirical power between LRST for continuous data and LRST-ordinal for discretized data across various sample sizes (N) in Scenario 1. The placebo-to-treatment allocation ratio was 2:3. Effect sizes were preset at 2.21 for ADAS-cog11 and 5.38 for DAD. Abbreviation: ADAS-cog11, Alzheimer's Disease Assessment Scale - cognitive subscale (11 items); DAD, Disability Assessment for Dementia.

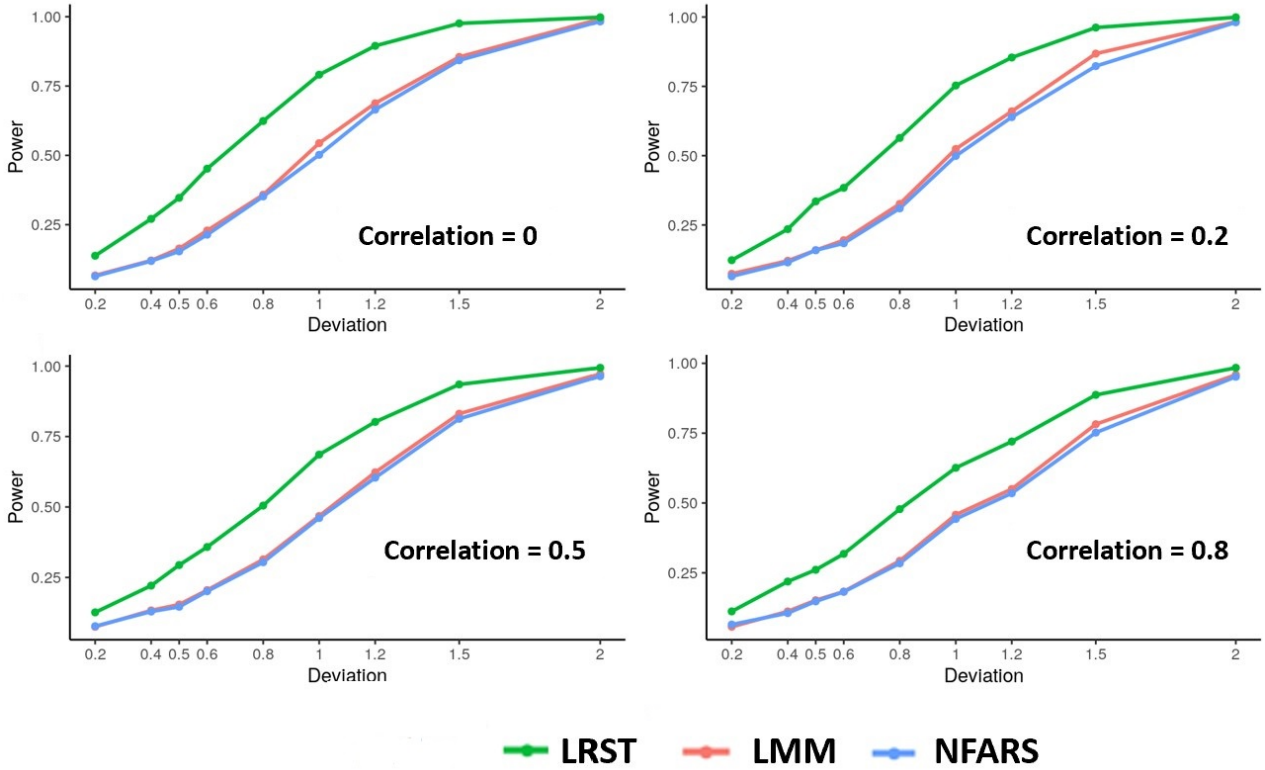

Figure 2: Empirical power curves of LRST, NFARS, and LMM methods under Scenario 2. This figure illustrates the power profiles for varying effect sizes, i.e., from 0.2 to 2 times those in Scenario 1 (labeled as "0.2 to 2 deviations"), across different inter-outcome correlation coefficients (0, 0.2, 0.5, and 0.8). The sample size is $N = 759$ with $n_x = 311$ in the placebo group and $n_y = 448$ in the Bapi group. Abbreviation: NFARS, the nonparametric factorial ANOVA rank statistics (Brunner *and others*, 2017); LMM, linear mixed model.

4 Real Data Application in Neurodegenerative Disorders

4.1 Analysis of the Bapineuzumab 302 trial

The Bapineuzumab (Bapi) 302 study (Salloway *and others*, 2014) is a double-blind, randomized, placebo-controlled, phase 3 RCT aimed at evaluating the efficacy of intravenous Bapi as compared with placebo in patients with mild-to-moderate AD dementia. Conducted at 170 sites in the United States from December 2007 through April 2012, the study employed a 2:3 allocation ratio to randomize participants into placebo and Bapi treatment groups. Assessments were conducted at baseline, subsequently every 13 weeks over 78 weeks ($t = 13, 26, 39, 52, 65, 78$ weeks with $T = 6$).

The primary endpoints of the study were the scores on the 11-item cognitive subscale of the Alzheimer's Disease Assessment Scale (ADAS-cog11) and the Disability Assessment for Dementia (DAD). The ADAS-cog11, a comprehensive tool consisting of 11 items, is designed to evaluate cognitive functions in AD patients across various domains including memory, language, praxis, and orientation. This is accomplished through a series of tasks ranging from word recall to object naming and command following. Scores on the ADAS-cog11 vary from 0 to 70, where higher scores denote more severe cognitive impairment, thus aiding clinicians in diagnosis and monitoring of disease progression. On the other hand, the DAD score, derived from caregiver assessments over 40 items, quantifies the patient's autonomy in daily tasks, such as personal hygiene and meal preparation. This scale ranges from 0 to 100, where a score of 100 reflects complete independence in all assessed activities, providing a metric to evaluate functional abilities and tailor interventions accordingly. The study's dataset comprised 759 participants, with 311 (n_x) in the placebo arm and 448 (n_y) receiving Bapi treatment.

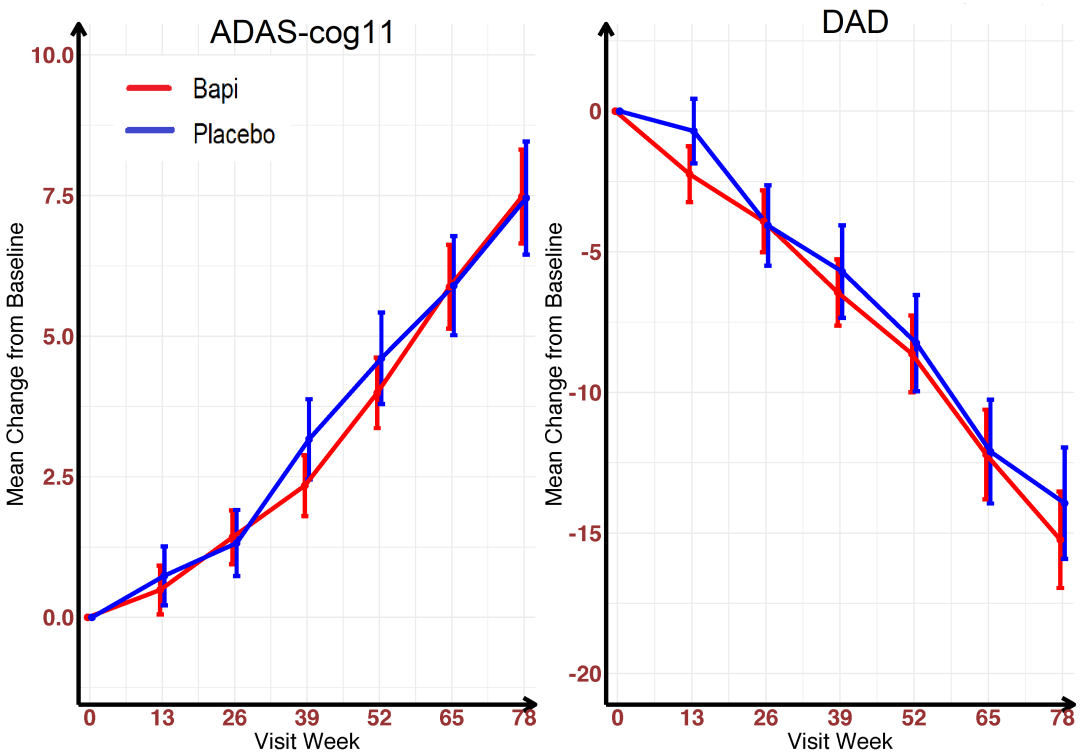

Figure 3: Mean changes from baseline to week 78 in scores on ADAS-cog11 (left panel, with scores ranging from 0 to 70 and higher scores indicating greater impairment) and DAD (right panel, with scores ranging from 0 to 100 and higher scores indicating less impairment) in the Bapineuzumab 302 Trial. Vertical lines are 95% confidence intervals. Abbreviation: ADAS-cog11, Alzheimer's Disease Assessment Scale - cognitive subscale (11 items); DAD, Disability Assessment for Dementia.

Figure 3 displays the mean changes from baseline for both ADAS-cog11 and DAD scores in the placebo and Bapi treatment groups. The increase in ADAS-cog11 scores over 78 weeks suggests an ongoing decline in cognitive function, while the concurrent decrease in DAD scores indicates a continuous loss of daily functioning capabilities in participants. These trends, indicative of the gradual deterioration characteristic of AD in both cognitive and functional domains, are observable in both treatment groups. However, no significant differences in the primary outcomes between the Bapi and placebo groups were observed. The LRST statistic was -0.0474 (SE: 1.950), resulting in a Z-statistics of -0.024 ($P = 0.509$), suggesting no significant differences in scores between the Bapi and the placebo groups across cognitive and functional domains.

In comparison, the nonparametric NFARS method gave p-values of 0.500 (ADAS-cog11) and 0.427 (DAD). As the minimum of these p-values (0.427) exceeds the Bonferroni-adjusted significance level of 0.05/2=0.025, we do not reject the null hypothesis of no treatment efficacy. Moreover, using the LMM method, the p-values derived from the F-test for the Bapi effect in the mixed model are 0.740 (ADAS-cog11) and 0.473 (DAD). After adjusting for multiple testing, none of these p-values are less than 0.05/2=0.025, leading to a failure to reject the null hypothesis of no treatment effect. Hence, all three analytical approaches yield results consistent with the findings of the Bapi RCT, indicating that Bapi did not improve clinical outcomes related to cognition and daily function in patients with AD (Salloway *and others*, 2014).

4.2 Analysis of the Azilect Study

The Azilect study (Hattori *and others*, 2019) is a Phase 3, randomized, double-blind, placebo-controlled clinical trial designed to evaluate the efficacy of rasagiline in Japanese patients with early-stage Parkinson's Disease (PD), the second most prevalent neurodegenerative disorder. This trial enrolled patients aged between 30 and 79 years, diagnosed with PD within the previous five years. Participants were assigned randomly in a 1:1 ratio to receive either 1 mg/day of rasagiline or a placebo for up to 26 weeks. Assessments occurred at baseline, followed by weeks 6, 10, 14, 20, and 26 ($T = 5$). The primary endpoint was the Movement Disorder Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS), a widely recognized scale for assessing Parkinsonian symptoms in both clinical and research contexts (Regnault *and others*, 2019). The MDS-UPDRS includes 65 items across four sections, using a 5-point Likert scale (0-4, with higher scores indicating greater severity): Part I, Non-Motor Aspects of Experiences of Daily Living (13 items); Part II, Motor Aspects of Experiences of Daily Living (13 items); Part III, Motor Examination (33 items); and Part IV, Motor Complications (6 items). Additional details on the MDS-UPDRS scale are available in (Goetz *and others*, 2008).

We applied our proposed statistical tests to the multiple outcomes derived from the MDS-UPDRS scale, focusing on the sum scores for Parts I (0 to 52), II (0 to 52), and III (0 to 132), thereby considering three primary outcomes ($K = 3$). Out of 209 patients ($N = 209$) who completed all study visits, 100 were in the placebo group ($n_x = 100$) and 109 received rasagiline ($n_y = 109$). Figure 4 illustrates the mean changes from baseline to week 26 in MDS-UPDRS Parts I, II, and III sum scores. Notably, the mean change in the MDS-UPDRS Part I sum score from baseline was similar across both groups. However, while the placebo group showed a consistent increase in the MDS-UPDRS Part II sum score change from baseline (indicating deterioration), the rasagiline group exhibited some improvements during the study, as compared to the baseline, although the improvements diminished towards the study's conclusion. The treatment effect of rasagiline on the MDS-UPDRS Part II sum score demonstrated a growing trend over the study period compared to placebo. For the MDS-UPDRS Part III sum score change from baseline, both groups showed steady improvement over the study duration, with the rasagiline group exhibiting more significant improvement and efficacy than the placebo group.

To align higher scores with clinically better outcomes, as defined in Section 2, we inverted all ordinal item scores (multiplied by -1) to calculate sum scores. The LRST statistic was 4.929 (SE: 1.572), resulting in a Z-statistic of 3.135 ($P = 8.592 \times 10^{-4}$), indicating a significant overall treatment effect of rasagiline on the primary outcomes of MDS-UPDRS Parts I, II, and III compared to placebo. Following the significant overall treatment effect, we fitted three univariate LMMs using MDS-UPDRS Parts I, II, and III sum scores as response variables. The p-values obtained from the F-test for the rasagiline effect in the mixed model are 0.418 (Part I), 2.95×10^{-4} (Part II), and 2.03×10^{-5} (Part III). Since the last two p-values are below the Bonferroni-adjusted significance level of 0.05/3=0.017, we concluded that rasagiline significantly improved motor aspects of daily living and motor examination in MDS-UPDRS Parts II and III, but did not effectively improve non-motor aspects of daily living in Part I.

Additionally, we applied nonparametric NFARS method to MDS-UPDRS Parts I, II, and III sum scores separately. The resulting p-values for these tests are 0.174, 0.0001, and 0.0013. Given that only the p-values for MDS-UPDRS Parts II and III sum scores were below the Bonferroni-adjusted significance level of 0.05/3=0.017, our conclusions are consistent with those from the LRST. Both the LRST and nonparametric NFARS corroborate the visible group differences in Parts II and III score changes (Figure 4, middle and lower panels), with minimal differences in Part I score changes (Figure 4, upper panel).

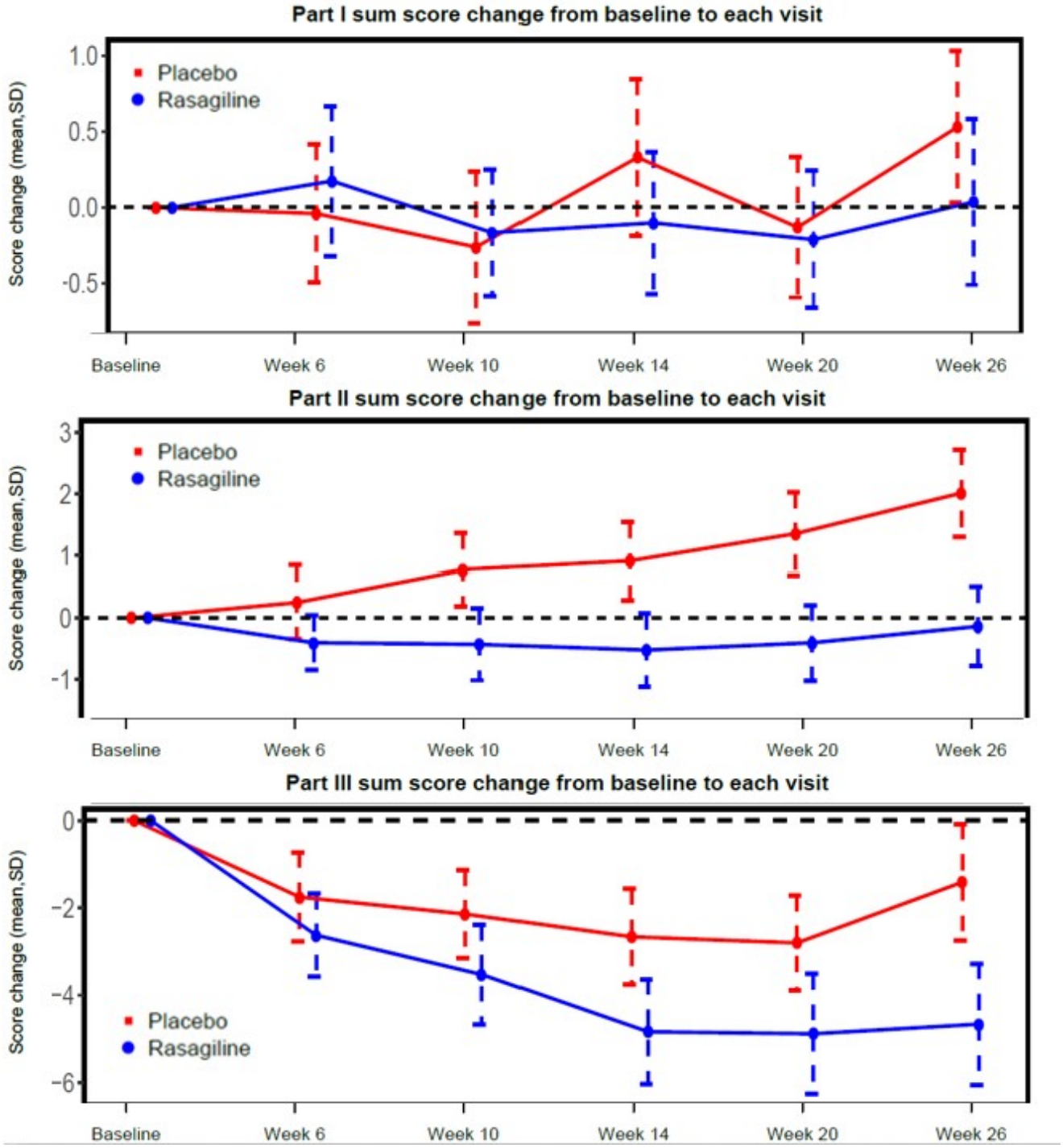

Figure 4: Mean changes from baseline to week 26 in MDS-UPDRS Parts I, II, and III sum scores. The vertical lines are the 95% confidence intervals. The dashed horizontal line indicates the baseline score reference. Dotted and dashed vertical lines represent standard deviations. Abbreviations: SD, standard deviation; MDS-UPDRS, Movement Disorder Society Unified Parkinson's Disease Rating Scale.

5 Discussions

This article develops and validates the longitudinal Rank Sum Test (LRST), a novel nonparametric rank-based hypothesis-testing statistic, for analyzing multiple longitudinal primary endpoints in randomized controlled trials of neurodegenerative disorders such as AD and PD. The LRST provides a robust and efficient statistical framework for assessing treatment efficacy and capturing dynamic changes in multiple endpoints over time. Our simulation study revealed that the LRST significantly outperforms traditional univariate tests requiring multiple comparison adjustments, such as Linear Mixed Models (LMM) and nonparametric factorial ANOVA rank statistics, by achieving higher statistical power in identifying overall treatment effects while precisely controlling Type I error rates. Applied to data from the Bapineuzumab AD study and the Azilect PD study, the LRST effectively detected overall treatment benefits, aligning with the primary conclusions of these studies. Following the determination of overall treatment efficacy, methodologies like LMM or Generalized Estimating Equations (GEE) can be employed to identify and quantify the treatment effects on individual outcomes.

The LRST statistic assigns equal weight across all time points and outcomes. To enhance the analysis's clinical relevance, we introduce a weighting mechanism that assigns variable weights over time, e.g., larger weights to in-clinic over telemedicine visits. To facilitate this, we may define a weighted sum of rank differences between groups across time points, denoted by $\boldsymbol{R^w}=(w_1(\bar{R}_{y\cdot 1\cdot}-\bar{R}_{x\cdot 1\cdot}),\cdots,w_T(\bar{R}_{y\cdot T\cdot}-\bar{R}_{x\cdot T\cdot}))'$, where the weight $w_t \geq 0$ and $\sum_{t=1}^{T} w_t = 1$. We then follow the procedure in Section 2 to develop a weighted LRST. This weighting methodology allows the LRST to more accurately capture the clinical significance of each visit type and can be applied to various outcomes. The weighted LRST simplifies to the original LRST when weights are equal. Moreover, optimizing weights via the covariance estimator $\hat{\boldsymbol{\Sigma}}$ may lower the weighted LRST's asymptotic variance, consistent with previous findings (Wei and Johnson, 1985; Wei *and others*, 1989). Development of the weighted LRST represents a future research direction.

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 LRST can be applied. Under the weighting paradigm, the kernel function $\boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{y}_j)$ (defined in Section S1 in the Supplementary material) of the U-statistic needs to be weighted by the estimated inverse probabilities to account for the missing outcome data. The development of consistent estimators for response probabilities and their integration into the longitudinal rank-sum test procedure remains an area for future research.

6 Software

The R package of LRST and implementation codes are available at: https://github.com/djghosh1123/LRST.

Acknowledgement

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

This study, carried out under YODA Project #2020-4448, used data of the Bapineuzumab 302 study obtained from the Yale University Open Data Access Project, which has an agreement with JANSSEN RESEARCH & DEVELOPMENT, L.L.C.. The interpretation and reporting of research using this data are solely the responsibility of the authors and do not necessarily represent the official views of the Yale University Open Data Access Project or JANSSEN RESEARCH & DEVELOPMENT, L.L.C..

The Azilect study data 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.

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

Akritas, MG and Brunner, E. (1997). A unified approach to rank tests for mixed models. *Journal of Statistical Planning and Inference* **61**, 249–277.

Association, Alzheimer's. (2023). 2023 Alzheimer's disease facts and figures. *Alzheimer's & Dementia* **19**(4), 1598–1695.

Bartholomew, DJ. (1961). A test of homogeneity of means under restricted alternatives. *Journal of the Royal Statistical Society, Series B* **23**(1), 239–272.

Bittman, R, Romano, JP, Vallarino, C and Wolf, M. (2009). Optimal testing of multiple hypotheses with common effect direction. *Biometrika* **96**(2), 399–410.

Brunner, E, Domhof, S and Langer, F. (2002*a*). *Nonparametric Analysis of Longitudinal Data in Factorial Experiments*. New York: Wiley.

Brunner, E, Konietschke, F, Pauly, M and ML, Puri. (2017). Rank-based procedures in factorial designs: hypotheses about non-parametric treatment effects. *Journal of the Royal Statistical Society Series B (Statistical Methodology)* **79**(5), 1463–1485.

Brunner, E, Munzel, U and Puri, ML. (2002*b*). The multivariate nonparametric Behrens–Fisher problem. *Journal of Statistical Planning and Inference* **108**(1-2), 37–53.

Department of Health & Human Services, U.S. *and others*. (2018). Early Alzheimer's Disease: Developing Drugs For Treatment, Guidelines for Industry.

Dobler, D, Friedrich, S and Pauly, M. (2020). Nonparametric MANOVA in meaningful effects. *Annals of the Institute of Statistical Mathematics* **72**, 997–1022.

Doody, RS, Raman, R, Farlow, M, Iwatsubo, T, Vellas, B, Joffe, S, Kieburtz, K, He, F, Sun, X, Thomas, RG *and others*. (2013). A phase 3 trial of semagacestat for treatment of Alzheimer's disease. *New England Journal of Medicine* **369**(4), 341–350.

Doody, RS, Thomas, RG, Farlow, M, Iwatsubo, T, Vellas, B, Joffe, S, Kieburtz, K, Raman, R, Sun, X, Aisen, PS *and others*. (2014). Phase 3 trials of solanezumab for mild-to-moderate Alzheimer's disease. *New England Journal of Medicine* **370**(4), 311–321.

Gauthier, S, Feldman, HH, Schneider, LS, Wilcock, GK, Frisoni, GB, Hardlund, JH, Moebius, HJ, Bentham, P, Kook, KA, Wischik, DJ *and others*. (2016). Efficacy and safety of tau-aggregation inhibitor therapy in patients with mild or moderate Alzheimer's disease: a randomised, controlled, double-blind, parallel-arm, phase 3 trial. *The Lancet* **388**(10062), 2873–2884.

Goetz, CG, Tilley, BC, Shaftman, SR, Stebbins, GT, Fahn, S, Martinez-Martin, P, Poewe, W, Sampaio, C, Stern, MB, Dodel, R, Dubois, B, Holloway, R, Jankovic, J, Kulisevsky, J, Lang, AE, Lees, A, Leurgans, S, LeWitt, PA, Nyenhuis, D, Olanow, CW, Rascol, O, Schrag, A, Teresi, JA, Hilten, JJ van *and others*. (2008). Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. *Movement Disorders* **23**(15), 2129–2170.

GUIDANCE, DRAFT. (2022). Multiple endpoints in clinical trials guidance for industry. *Center for Biologics Evaluation and Research (CBER)*.

Gunawardana, A and Konietschke, F. (2019). Nonparametric multiple contrast tests for general multivariate factorial designs. *Journal of Multivariate Analysis* **173**, 165–180.

Hamasaki, T, Evans, SR and Asakura, K. (2018). Design, data monitoring, and analysis of clinical trials with co-primary endpoints: a review. *Journal of Biopharmaceutical Statistics* **28**(1), 28–51.

Hansson, O, Blennow, K, Zetterberg, H and Dage, J. (2023). Blood biomarkers for Alzheimer's disease in clinical practice and trials. *Nature Aging* **3**(5), 506–519.

Hattori, N, Takeda, A, Takeda, S, Nishimura, A, Kitagawa, T, Mochizuki, H, Nagai, M and Takahashi, R. (2019). Rasagiline monotherapy in early Parkinson's disease: A phase 3, randomized study in Japan. *Parkinsonism and Related Disorders* **60**, 146–152.

Huang, P, Tilley, BC, Woolson, RF and Lipsitz, S. (2005). Adjusting O'Brien's test to control type 1 error for the generalized nonparametric Behrens-Fisher problem. *Biometrics* **61**(2), 532–539.

Konietschke, Frank, Bathke, Arne C, Hothorn, Ludwig A and Brunner, Edgar. (2010). Testing and estimation of purely nonparametric effects in repeated measures designs. *Computational Statistics & Data Analysis* **54**(8), 1895–1905.

Lachos, VH, Bandyopadhyay, D and Dey, DK. (2011). Linear and nonlinear mixed-effects models for censored HIV viral loads using normal/independent distributions. *Biometrics* **67**(4), 1594–1604.

Larocque, Denis, Haataja, Riina, Nevalainen, Jaakko and Oja, Hannu. (2010). Two sample tests for the nonparametric Behrens–Fisher problem with clustered data. *Journal of Nonparametric Statistics* **22**(6), 755–771.

Liu, A, Li, Q, Liu, C, Yu, K and Yu, KF. (2010). A rank-based test for comparison of multidimensional outcomes. *Journal of the American Statistical Association* **105**(490), 578–587.

Lu, ZH. (2013). Halfline tests for multivariate one-sided alternatives. *Computational Statistics and Data Analysis* **57**, 479–490.

Meng, Zhen, Li, Na and Yuan, Ao. (2022). Testing high-dimensional nonparametric Behrens-Fisher problem. *Journal of Systems Science and Complexity* **35**(3), 1098–1115.

Mintzer, J, Lanctôt, KL, Scherer, RW, Rosenberg, PB, Herrmann, N, Van Dyck, CH, Padala, PR, Brawman-Mintzer, O, Porsteinsson, AP, Lerner, AJ *and others*. (2021). Effect of methylphenidate on apathy in patients with Alzheimer disease: the ADMET 2 randomized clinical trial. *JAMA Neurology* **78**(11), 1324–1332.

Munzel, Ullrich and Brunner, Edgar. (2000). Nonparametric methods in multivariate factorial designs. *Journal of Statistical Planning and Inference* **88**(1), 117–132.

Noguchi, K, Gel, YR, Brunner, E and Konietschke, F. (2012). nparLD: an R software package for the nonparametric analysis of longitudinal data in factorial experiments. *Journal of Statistical Software* **50**(12), 1–23.

O'Brien, PC. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics* **40**(4), 1079–1087.

Perlman, MD. (1969). One-sided testing problems in multivariate analysis. *The Annals of Mathematical Statistics* **40**(2), 549–567.

Regnault, A, Boroojerdi, B, Meunier, J, Bani, M, Morel, T and Cano, S. (2019). Does the MDS-UPDRS provide the precision to assess progression in early Parkinson's disease? Learnings from the Parkinson's progression marker initiative cohort. *Journal of Neurology* **266**, 1927–1936.

Roy, A, Harrar, SW and Konietschke, F. (2019). The nonparametric Behrens-Fisher problem with dependent replicates. *Statistics in Medicine* **38**, 4939–4962.

Rubarth, K, Pauly, M and F, Konietschke. (2022*a*). Ranking procedures for repeated measures designs with missing data: Estimation, testing and asymptotic theory. *Statistical Methods in Medical Research* **31**(1), 105–118.

Rubarth, K, Sattler, P, Zimmermann, HG and Konietschke, F. (2022*b*). Estimation and testing of Wilcoxon-Mann-Whitney effects in factorial clustered data designs. *Symmetry* **14**(244), 1–34.

Salloway, S, Sperling, R, Fox, NC, Blennow, K, Klunk, W, Raskind, M, Sabbagh, M, Honig, LS, Porsteinsson, AP, Ferris, S *and others*. (2014). Two phase 3 trials of bapineuzumab in mild-to-moderate Alzheimer's disease. *New England Journal of Medicine* **370**(4), 322–333.

Scheiblich, H, Trombly, M, Ramirez, A and Heneka, MT. (2020). Neuroimmune connections in aging and neurodegenerative diseases. *Trends in Immunology* **41**(4), 300–312.

Tang, DI. (1994). Uniformly more powerful tests in a one-sided multivariate problem. *Journal of the American Statistical Association* **89**, 1006–1011.

Tang, DI, Geller, NL and Pocock, SJ. (1993). On the design and analysis of randomized clinical trials with multiple endpoints. *Biometrics* **49**(1), 23–30.

Umlauft, M, Placzek, M, F, Konietschke. and Pauly, M. (2019). Wild bootstrapping rank-based procedures: Multiple testing in nonparametric factorial repeated measures designs. *Journal of Multivariate Analysis* **171**, 176–192.

Wei, LJ and Johnson, WE. (1985). Combining dependent tests with incomplete repeated measurements. *Biometrika* **72**(2), 359–364.

Wei, LJ, Lin, DY and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association* **84**(408), 1065–1073.

Yu, JC, Hlávka, JP, Joe, E, Richmond, FJ and Lakdawalla, DN. (2022). Impact of non-binding FDA guidances on primary endpoint selection in Alzheimer's disease trials. *Alzheimer's & Dementia: Translational Research & Clinical Interventions* **8**(1), e12280.

Zhang, W, Liu, A, Tang, LL and Li, Q. (2019). A cluster-adjusted rank-based test for a clinical trial concerning multiple endpoints with application to dietary intervention assessment. *Biometrics* **75**(3), 821–830.

Zhu, H, Chen, K, Luo, X, Yuan, Y and Wang, JL. (2019). Fmem: Functional mixed effects models for longitudinal functional responses. *Statistica Sinica* **29**(4), 2007–33.

Zhuang, Y, Guan, Y, Qiu, L, Lai, M, Tan, MT and Chen, P. (2018). A novel rank-based non-parametric method for longitudinal ordinal data. *Statistical Methods in Medical Research* **27**(9), 2775–2794.