Asian Journal of Probability and Statistics

Volume 26, Issue 2, Page 62-75, 2024; Article no.AJPAS.112095

ISSN: 2582-0230

Multivariate Rank-Based Analysis of Multiple Endpoints in Clinical Trials: A Global Test Approach

Kexuan Li [a*], Lingli Yang [b], Shaofei Zhao [c], Susie Sinks [d], Luan Lin [d] and Peng Sun [d]

[a] *Global Biometrics and Data Sciences Bristol Myers Squibb, Cambridge, Massachusetts, US.*

[b] *Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, Massachusetts, US.*

[c] *Data and Statistical Sciences, AbbVie, Madison, New Jersey, US.*

[d] *Global Analytics and Data Sciences, Biogen, Cambridge, Massachusetts, US.*

Authors' contributions

This work was carried out in collaboration among all authors. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/AJPAS/2024/v26i2592

Open Peer Review History:

This journal follows the Advanced Open Peer Review policy. Identity of the Reviewers, Editor(s) and additional Reviewers, peer review comments, different versions of the manuscript, comments of the editors, etc are available here: http://www.sdiarticle5.com/review-history/112095

Received: 02/12/2023

Accepted: 05/02/2024

Published: 12/02/2024

Original Research Article

Abstract

Clinical trials demand a comprehensive evaluation of multiple endpoints to provide a thorough understanding of intervention efficacy and safety. In response to a discerned gap in existing literature, this study introduces an innovative global nonparametric testing procedure, grounded in multivariate ranks, for the holistic analysis of these diverse endpoints. Unlike conventional approaches that heavily rely on pairwise comparisons for individual endpoints, our method takes a novel approach by directly incorporating multivariate ranks. This methodology builds on the strengths of previous models while addressing their limitations, ensuring a more

**Corresponding author: E-mail: kexuan.li.77@gmail.com;*

nuanced and robust assessment. By considering the joint ranking of all endpoints, our approach exhibits heightened resilience against diverse data distributions and common censoring mechanisms commonly encountered in clinical trials. The proposed method emerges from a thoughtful integration of existing models, contributing to the methodological evolution in the field. To illustrate its superiority, extensive simulations have been conducted. The results unequivocally demonstrate the superior performance of our multivariate rank-based approach, showcasing its ability to effectively control type I error and achieve higher power compared to conventional rank-based methods. This empirical validation not only underscores the method's efficacy but also highlights its versatility and robustness across a spectrum of clinical trial settings.

Keywords: Global testing; multivariate rank; multiple endpoints; rank-based methods.

2010 Mathematics Subject Classification: 53C25; 83C05; 57N16

1 Introduction

In clinical trials, patients are often evaluated using multiple measures of treatment effectiveness, such as survival time, biomarker dynamics, and functional evaluations. To ensure a comprehensive evaluation of therapeutic benefits, it is important to consider multiple endpoints simultaneously. For instance, spinal muscular atrophy (SMA), a rare neuromuscular disorder causing motor neuron loss and progressive muscle degeneration, is assessed in treatment trials by comparing therapies based on both survival time and changes in muscle function measured using the Hammersmith Functional Motor Scale-Expanded (HFMSE). To draw a conclusive judgment on the effectiveness of treatment across multiple outcomes, several approaches have been suggested to summarize the treatment effect. One such approach involves employing multiple testing correction techniques, which adjust the individual p-values obtained from statistical tests to account for the possibility of false positives. Examples of such correction methods include the Bonferroni correction [1] and the Benjamini-Hochberg procedure [2]. Nevertheless, these methods are unable to provide a comprehensive assessment regarding the overall efficacy of clinical intervention. Furthermore, the endpoints measured in a clinical trial often exhibit high correlation, which introduces further challenges when utilizing these approaches. Often, however, people might be more interested in highlighting whether a subset of variables is jointly suggestive of a treatment effect [3]. In such cases, instead of testing each outcome individually, an overall statement should be claimed to assess therapeutic benefit. An alternative approach to address the issue of multiplicity is to create a composite endpoint by merging all relevant clinical outcomes into a single variable. This allows for an evaluation of the overall therapeutic benefit in a comprehensive manner, providing a "global" assessment. By conducting a single statistical test on the composite endpoint, there is no need for adjustment or correction for multiple comparisons. This approach simplifies the analysis and interpretation process by considering all relevant outcomes simultaneously. In essence, the global test procedure transforms a multivariate problem into a univariate scale, enabling the announcement of a single probability statement regarding the success of targeted intervention strategies. By utilizing a global testing procedure, it becomes possible to summarize the overall effect of treatment across multiple outcomes in a more flexible manner. This approach allows for a comprehensive evaluation of treatment efficacy, considering the collective impact of various outcomes simultaneously. In this paper, we introduce a nonparametric global testing procedure based on (multivariate) rank energy statistics developed by [4]. This procedure is designed to summarize the overall treatment effect across multiple measurements, including censored outcomes. The proposed method enables the integration of various measurements, accommodating both continuous, discrete, and censored data, to provide a comprehensive assessment of the treatment's impact across multiple measurements, including censored outcomes.

Before proceeding, let us first review some of the global test procedures that have been proposed in the literature. To name a few, [5] proposed a nonparametric rank-sum-type test to compare the distribution of two samples with multiple outcomes by summing up the ranks for each individual outcome and the test statistics is proven to be asymptotically normal distributed under the null hypothesis that the two multivariate samples have the same distribution. [6] presented a nonparametric test for time to event endpoint and longitudinal measure. [7] introduced a family of simple testing procedures for scoring multivariate ordinal data. [8] considered the sample size computation problem for clinical trial design with multiple primary outcomes. [9] constructed a hierarchical global ranking of a survival endpoint and a longitudinal measure to test the null hypothesis that neither of the outcomes is associated with treatment. [10] described a new endpoint for amyotrophic lateral sclerosis by combining survival time and change in function score. [11] further generalized the aforementioned global nonparametric rank tests using U-statistics and discussed the choice of optimal weighting schemes. Specifically, Let x_{ik}, y_{jk} be the observed outcome k for subject i in the control group, $i = 1, \ldots, m$ and subject j in the treatment group, $j = 1, \ldots, n$, where $k = 1, \ldots, d$. Denote $x_i = (x_{i1}, \ldots, x_{id})^\top, y_j = (y_{j1}, \ldots, y_{jd})^\top$ as the observed vector. [11] defined $r_{ij}^{(k)}(x_{ik}, y_{jk}) : \mathbb{R} \times \mathbb{R} \to [-1, 1]$ be the rank score between i-th subject in control group and j-th subject in the treatment group for outcome k. For example, $r_{ij}^{(k)}(x_{ik}, y_{jk}) = 1(x_{ik} > y_{jk}) - 1(x_{ik} < y_{jk})$. [11] further defined $r_{ij}(x_i, y_j) = (r_{ij}^{(1)}(x_{i1}, y_{j1}), \ldots, r_{ij}^{(d)}(x_{id}, y_{jd}))$ as the vector of rank scores between i-th subject and j-th subject, and for simplicity, we sometimes write r_{ij} for $r_{ij}(x_i, y_j)$, $r_{ij}^{(k)}$ for $r_{ij}^{(k)}(x_{ik}, y_{jk})$ if no confusion arises. Then, the two-sample U-statistics generalized by [11] was defined as the summation among all the pairwise comparisons between two groups after mapping r_{ij} to a univariate score:

$$U = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \phi(r_{ij}), \tag{1.1}$$

where $\phi : \mathbb{R}^d \to \mathbb{R}$ maps the vector of comparison for each outcome to a one-dimensional score, which gives us the overall evaluation. Ideally, the maximizer and minimizer of $\phi(r_{ij})$ should be $(1, \ldots, 1)$ and $(-1, \ldots, -1)$, respectively. Several choices of ϕ have been derived in literature. For example, in [5], the author proposed a nonparametric procedure that calculated an overall rank by summing outcome-specific ranks for each subject and used a two-sample rank-sum or t test to test for the null hypothesis. In other word, map ϕ in [5] is given by

$$\phi(r_{ij}) = r_{ij}^{(1)} + r_{ij}^{(2)} + \ldots + r_{ij}^{(d)}. \tag{1.2}$$

In [5], the underlying assumption is that each outcome has the same importance. However, in practice, not all the endpoints may contribute equally to the treatment effect, so, the rank-sum test by [5] could be further generalized to the weighted summation as $\phi(r_{ij}) = \sum_{k=1}^{p} w_k r_{ij}^{(k)}$, where $w_k \geq 0$ is the weight associated with each component. The Finkelstein-Schoenfeld test introduced by [6] compared a mortality outcome and a longitudinal outcome in a hierarchical manner, where subjects are first compared pairwise on survival using the Gehan scoring function [12], and then on the longitudinal marker if it is indeterminate who survived longer. In their framework, the function ϕ is defined as

$$\phi(r_{ij}) = r_{ij}^{(1)} + 1(r_{ij}^{(1)} = 0) r_{ij}^{(2)} + \ldots + 1(r_{ij}^{(1)} = \ldots = r_{ij}^{(d-1)} = 0) r_{ij}^{(d)}.$$

In [7], the authors conducted pairwise comparisons of subjects based on ordinal measures. They considered two situations: (1) assigning a score of 1 to the subject whose outcomes are all favorable, and (2) assigning a score of 1 to the subject who has more favorable outcomes. The corresponding function ϕ can be expressed as

$$\phi(r_{ij}) = 1(\max_{k=1,\ldots,d} \{r_{ij}^{(k)}\} > 0) - 1(\min_{k=1,\ldots,d} \{r_{ij}^{(k)}\} < 0), \text{ for situation (1)}$$

$$\phi(r_{ij}) = 1(\sum_{k=1}^{d} r_{ij}^{(k)} > 0) - 1(\sum_{k=1}^{d} r_{ij}^{(k)} < 0), \text{ for situation (2).}$$

All the methods mentioned above utilized the univariate rank of each individual endpoint and mapped each pair to a one-dimensional score. In addition to rank-based testing procedures, various other testing procedures have been explored in the literature for handling multiple endpoints, such as [13, 14, 15]. For a comprehensive review of these and other testing procedures, please refer to [3].

The remainder of the paper is organized as follows. Section 2 formulates the statistical problem and presents nonparametric global rank testing for survival and multiple endpoints using multivariate ranks as well as the implementation. Section 3 evaluates the finite-sample performance of the proposed methodology by several simulation studies and some concluding remarks are provided in Section 4.

2 Methodology

Before presenting the proposed nonparametric global test statistics, we first formulate the problem under study. Suppose in a clinical trial, people collect d measurements to verify the efficacy of an intervention. Let x_{ik}, y_{jk} be the observed outcome k for subject i in the control group, $i = 1, \dots, m$ and subject j in the treatment group, $j = 1, \dots, n$, where $k = 1, \dots, d$. Denote $x_i = (x_{i1}, \dots, x_{id})^\top, y_j = (y_{j1}, \dots, y_{jd})^\top$ as the observed vector. Suppose $x_1, \dots, x_m \overset{\text{i.i.d.}}{\sim} \mu_x$ and $y_1, \dots, y_n \overset{\text{i.i.d.}}{\sim} \mu_y$, where μ_x, μ_y are two d-dimensional distributions. The following nonparametric two-sample goodness-of-fit testing problem is considered

$$H_0 : \mu_x = \mu_y, \text{ versus } H_1 : \mu_x \neq \mu_y. \tag{2.1}$$

When $d = 1$ and μ_x, μ_y are unknown, the problem is a classical nonparametric two-sample test and several methods have been proposed for it, which includes Spearman's rank correlation test [16], two-sample Cramér-von Mises statistic [17], two-sample Kolmogorov-Smirnov test [18], Wald-Wolfowitz run test [19], Wilcoxon-Mann-Whitney rank test [20], Hoeffding's D-test [21], among others. In the case when $d > 1$, the problem of nonparametric two-sample testing for multivariate distributions has also a long history and has recently gained significant attention. Various methods have been proposed, such as those by [22, 23, 24, 25, 26, 27, 28, 29], [30], [31], [32]. In this work, we specifically focus on the (multivariate) rank-based approach proposed by [4].

2.1 Multivariate Rank

[4] introduced the concept of multivariate rank, which utilizes low-discrepancy sequences to transform the original data into a unit hypercube. Without loss of generality, we let $\mathbb{R}^d$ be the d-dimensional input space, and the d-dimensional unit hypercube to which the data is mapped by the multivariate rank process is represented by $[0, 1]^d$. The families of all probability distributions on $\mathbb{R}^d$ are denoted by $\mathcal{P}(\mathbb{R}^d)$, while the families of Lebesgue absolutely continuous probability measures on $\mathbb{R}^d$ are represented by $\mathcal{P}_{ac}(\mathbb{R}^d)$, and the uniform distribution on $[0, 1]^d$ is denoted by $\mathcal{U}^d$. The multivariate rank approach is implemented using a measure transportation technique, also known as optimal transportation. Specifically, this involves finding a suitable function $G : \mathbb{R}^d \to \mathbb{R}^d$ that maps a given measure $\mu \in \mathcal{P}(\mathbb{R}^d)$ to $\nu \in \mathcal{P}(\mathbb{R}^d)$, represented as $G\#\mu = \nu$. In other words, if x follows the distribution μ, then $G(x)$ follows the distribution ν. Suppose the observed data follows distribution $\mu \in \mathcal{P}_{ac}(\mathbb{R}^d)$, then the idea of multivariate rank introduced by [4] is to find a rank function $R(\cdot)$ such that $R\#\mu = \mathcal{U}^d$, that is, $R(x)$ follows a uniform distribution in $\mathbb{R}^d$. The following theorem guarantees the existence of the (population) rank function $R(\cdot)$.

Theorem 2.1 (McCann's theorem [33])**.** *Suppose $\mu, \nu \in \mathcal{P}_{ac}(\mathbb{R}^d)$, then there exists a unique function $R(\cdot)$, up to measure zero sets with respect to μ, which is the gradient of an extended real-valued d-variate convex function, such that $R\#\mu = \nu$. Moreover, if μ and ν have finite second moments, then $R(\cdot)$ is also the solution to Monge's problem, i.e.,*

$$R(\cdot) =_G \int \|x - G(x)\|^2 d\mu(x), \quad \textit{subject to } G\#\mu = \nu. \tag{2.2}$$

However, in practice, it is infeasible to know the true distribution μ, instead, the only available information regarding the measure μ is obtained through empirical observations $x_1, \dots, x_n$. [4] put forward a novel approach for estimating the (population) rank function using empirical observations. To understand this approach, it is first necessary to review the definition of the low-discrepancy sequence, which plays a crucial role in the method.

A low-discrepancy sequence is a sequence of points in a multi-dimensional space that is designed to have better distribution properties than random sequences. In particular, these sequences are constructed to have a low discrepancy, which measures the uniformity of the distribution of points in a given region. To be more precise, let us consider a d-dimensional hypercube $[0, 1]^d$ and let $\mathcal{A}$ be a set of n points in this hypercube. The discrepancy of a set $\mathcal{A}$ is defined as:

$$D(\mathcal{A}) = \sup_{\mathcal{B} \subset [0,1]^d} \left| \frac{\#(\mathcal{A} \cap \mathcal{B})}{n} - \text{Leb}(\mathcal{B}) \right|,$$

where $\mathcal{B}$ is any region in $[0,1]^d$, $\#(\mathcal{A} \cap \mathcal{B})$ denotes the number of points in $\mathcal{A}$ that fall in $\mathcal{B}$, and Leb($\mathcal{B}$) is the Lebesgue measure of $\mathcal{B}$. In other words, the discrepancy measures the maximum difference between the fraction of points in $\mathcal{A}$ that fall in any subinterval of $[0,1]^d$ and the measure of that subinterval. A low-discrepancy sequence is a sequence of points in $[0,1]^d$ that has a small discrepancy and the proportion of points in the sequence falling into an arbitrary set $\mathcal{B}$ is close to the proportional of the measure of $\mathcal{B}$. There are many different constructions of low-discrepancy sequences, such as Hammersley sequences [34], Halton sequences [35], and Sobol sequences [36]. [4] proposed a novel approach for estimating the population rank function using a low-discrepancy sequence of points. In their method, they first generated a low-discrepancy sequence of points in the d-dimensional space of features, which was a good representation of $\mathcal{U}^d$. Then, the (empirical) rank map was defined as the solution of the empirical version of Monge's problem in (2.2). To be more specific, let $x_1, ..., x_n \in \mathbb{R}^d$ be the $i.i.d.$ observations, and $\{c_1, ..., c_n\} \subset [0,1]^d$ be a low-discrepancy sequence on $[0,1]^d$. Let δ_a denote the Dirac measure that assigns probability 1 to the point a and $\mu_n^x(x) = n^{-1} \sum_{i=1}^n \delta_{x_i}, \nu_n = n^{-1} \sum_{i=1}^n \delta_{c_i}$. Then the (empirical) rank map is defined as

$$\widehat{R}(\cdot) =_F \int \|x - F(x)\|^2 d\mu_n^x(x), \quad \text{subject to } F\#\mu_n^x = \nu_n,$$

which is equivalent to the following optimization problem:

$$\widehat{\sigma} =_{\sigma=(\sigma(1),\ldots,\sigma(n))\in S_n} \sum_{i=1}^{n} \| \; x_i - c_{\sigma(i)} \; \|^2 =_{\sigma=(\sigma(1),\ldots,\sigma(n))\in S_n} \sum_{i=1}^{n} \langle x_i, c_{\sigma(i)} \rangle, \tag{2.3}$$

where $\| \cdot \|$ and $\langle \cdot, \cdot \rangle$ denote the usual Euclidean norm and inner product, and S_n is the set of all permutations of $\{1, 2, ..., n\}$. Finally, the (empirical) multivariate rank of x_i is a d-dimensional vector defined as

$$\widehat{R}(x_i) = c_{\widehat{\sigma}(i)}, \quad \text{for } i = 1, \ldots, m. \tag{2.4}$$

Fig. 1. illustrates the idea of univariate rank on $[0,1]$ and multivariate rank on $[0,1]^2$. In the following sections. we proceed to show how to construct a nonparametric global test for different types of endpoints based on multivariate rank. It is worthwhile to mention that the multivariate rank defined by [4] has also been extended to other areas, such as deep learning [37], and variable selection [38].

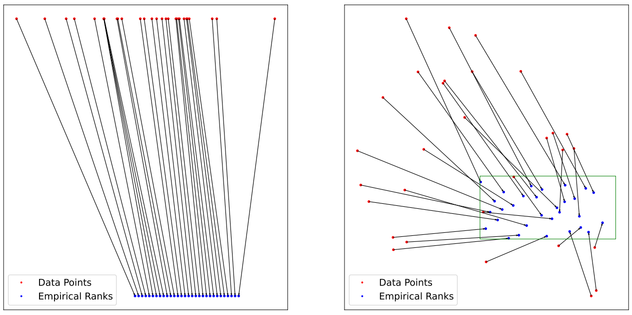

Fig. 1. Left: Univariate rank on $[0,1]$, the red dots are randomly generated from a standard normal distribution, and the blue dots are evenly spaced on $[0,1]$. The leftmost (smallest) red dot will be assigned to the first blue dot, thus rank 1, and so on. Right: Multivariate rank on $[0,1]^2$, the blue dots are two-dimensional Sobol' sequence on $[0,1]^2$, while the green box represents the region $[0,1]^2$. The multivariate rank method will map the two-dimensional red dots to the blue dots, which correspond to their multivariate ranks.

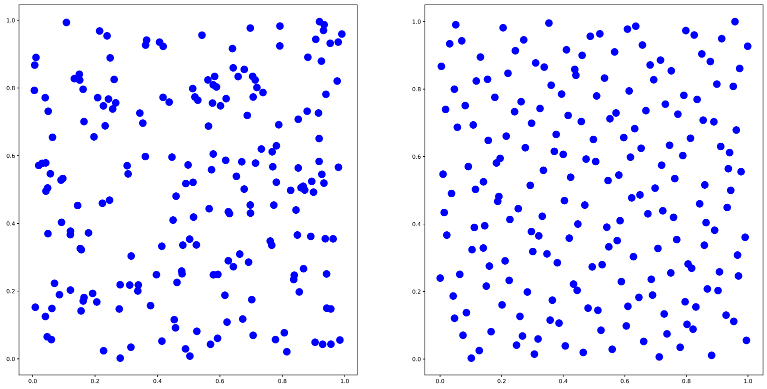

Fig. 2. Left: The dots are randomly generated from a uniform distribution. Right: The dots are generated using the Sobol' sequence. The Sobol' sequence provides a more even distribution of points compared to randomly generated points.

2.2 A Global Test Statistics Based on Multivariate Rank

We now proceed to study the nonparametric global test problem in (2.1) based on multivariate rank. Given the observations of control group $x_1, \ldots, x_m$ and treatment group $y_1, \ldots, y_n$, first, pool the sample of $(m+n)$ observations into a single group and get the (empirical) multivariate rank of x_i, y_j through (2.4), denoted as $\widehat{R}^{x,y}_{m,n}(x_i), \widehat{R}^{x,y}_{m,n}(y_j)$. The corresponding test statistic based on multivariate rank is defined as

$$\begin{aligned}\mathrm{RE}^2_{m,n} := \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left\|\widehat{R}^{x,y}_{m,n}(x_i) - \widehat{R}^{x,y}_{m,n}(y_j)\right\| - \frac{1}{m^2}\sum_{i,j=1}^{m}\left\|\widehat{R}^{x,y}_{m,n}(x_i) - \widehat{R}^{x,y}_{m,n}(x_j)\right\| \\ - \frac{1}{n^2}\sum_{i,j=1}^{n}\left\|\widehat{R}^{x,y}_{m,n}(y_i) - \widehat{R}^{y,y}_{m,n}(y_j)\right\|. \qquad (2.5)\end{aligned}$$

Under Theorem 4.3 in [4], we know $\mathrm{RE}^2_{m,n}$ is $O_p(1)$ with the following limiting distribution

$$\frac{mn}{m+n}\mathrm{RE}^2_{m,n} \xrightarrow{w} \sum_{j=1}^{\infty}\tau_j Z_j^2, \quad \min(m,n)\to\infty, \tag{2.6}$$

where Z_j's are *i.i.d.* standard normals and τ_j's are fixed nonnegative constants which does not depend on the distribution x_i, y_j, and $\xrightarrow{w}$ denotes the weak convergence of distributions. Given predetermined significance level α, let

$$c_{m,n} := \inf\left\{c > 0 : \mathbb{P}_{\mathrm{H}_0}\left(mn(m+n)^{-1}\mathrm{RE}^2_{m,n} \geq c\right) \leq \alpha\right\}, \tag{2.7}$$

and the decision rule for testing (2.1) at significance level α can be defined as follows

$$\phi_{m,n} = 1(mn(m+n)^{-1}\mathrm{RE}^2_{m,n} \geq c_{m,n}),$$

where $\mathrm{RE}^2_{m,n}, c_{m,n}$ are defined in (2.5) and (2.7), respectively. We reject the null hypothesis in (2.1) if and only if $\phi_{m,n} = 1$. By the definition of $c_{m,n}$, clearly, the test is level α. It is worth noting that for any fixed m, n, $\mathrm{RE}^2_{m,n}, c_{m,n}$ do not depend on μ_x, μ_y. Unfortunately, as mentioned in [4], the theoretical value of $c_{m,n}$ is infeasible to achieve, the only way to get $c_{m,n}$ as of now is through numerical experiments. We summarize the asymptotic thresholds for $mn(m+n)^{-1}\mathrm{RE}^2_{m,n}$ in Table 3. and provide the procedure to estimate $c_{m,n}$ in Algorithm 2.3.

Table 1. Asymptotic thresholds for $mn(m+n)^{-1}\mathrm{RE}^2_{m,n}$ when $\alpha = 0.05, 0.10, d \leq 6$. The numbers are obtained through Algorithm 2.3.

	$d=1$	$d=2$	$d=3$	$d=4$	$d=5$	$d=6$
$\alpha = 0.05$	0.94	1.12	1.26	1.37	1.45	1.54
$\alpha = 0.10$	0.70	0.92	1.07	1.17	1.28	1.37

```
Algorithm 2.2.  Input: $(x_1, \dots, x_m)$ and $(y_1, \dots, y_n)$.
  Pool $(x_1, \dots, x_m)$ and $(y_1, \dots, y_n)$ into a single group.
  Generate a low-discrepancy sequence $\{c_1, ..., c_{(m+n)}\}$ on $[0,1]^d$ with size $(m+n)$.
  Solve the optimal assignment problem:
      $\widehat{\sigma} =_{\sigma=(\sigma(1),\dots,\sigma(m+n))\in S_{m+n}} \left(\sum_{i=1}^{m} \| x_i - c_{\sigma(i)} \|^2 + \sum_{j=m+1}^{m+n} \| y_{j-m} - c_{\sigma(j)} \|^2\right).$
  Compute the (empirical) multivariate rank of $x_i, y_j$ by
      $\widehat{R}^{x,y}_{m,n}(x_i) = c_{\widehat{\sigma}(i)}, i = 1, \dots, m, \widehat{R}^{x,y}_{m,n}(y_j) = c_{\widehat{\sigma}(j+m)}, j = 1, \dots, n.$
  Return: test statistic $\mathrm{RE}^2_{m,n}$ in (2.5).
```

```
Algorithm 2.3.  Initialization: set $n_{run} := 10^6$ and $A = [0, \dots, 0]$ as a zero array with length $n_{run}$;
  Input: $m, n > 0, \alpha, p$;
  for $i$ in $1, \dots, n_{run}$ do
      generate $x_1, \dots, x_m, y_1, \dots, y_n$ from standard d-dimensional multivariate normal distribution independently;
      compute test statistics $\mathrm{RE}^2_{m,n}$ using Algorithm 2.2;
      $A[i] \leftarrow \mathrm{RE}^2_{m,n}$;
  end for
  Return: $c_{m,n} = (1-\alpha)$ quantile of $A$.
```

2.3 Time-to-Event Endpoint

In the following subsection, we consider the problem that one of the endpoint is the time-to-event outcome. Without loss of generality, we assume the first component of the d measurements is the survival endpoint and the rest $d-1$ components are non-survival endpoints. we assume x_{i1}, y_{j1} be the survival endpoint and $x_{i1} = \min\{T^x_i, C^x_i\}, y_{j1} = \min\{T^y_j, C^y_j\}, \delta^x_i = 1(T^x_i \leq C^x_i), \delta^y_j = 1(T^y_j \leq C^y_j)$, where $T^x_i(T^y_j), C^x_i(C^y_j), \delta^x_i(\delta^y_j)$ is the unknown survival time, censoring time, and the censoring indicator of subject $i(j)$ in control(treatment) group. In order to apply the global test based on multivariate rank for the survival endpoint, we use the idea from the Gehan–Wilcoxon test [12], which is an extension of the classical Wilcoxon rank-sum test for comparing survival curves between two or more groups. More specifically, in the first step, we pool the two samples of $(m+n)$ survival times $(x_{11}, \dots, x_{m1}, y_{11}, \dots, y_{n1})$ into a single group $(t_1, \dots, t_{m+n})$, and we use a superscript '+' to indicate that the corresponding observation is censored. Then we construct a score by comparing each individual with the remaining $(m+n-1)$ subjects based on the following rule:

$$\widetilde{u}_{ij} = \begin{cases} +1 & \text{if} \quad t_i > t_j \text{ or } t_i^+ \geq t_j, \\ -1 & \text{if} \quad t_i < t_j \text{ or } t_i \leq t_j^+, \\ 0 & \text{otherwise}. \end{cases} \tag{2.8}$$

Then the importance score for each individual is defined as $u_i = \sum_{j=1}^{m+n} \widetilde{u}_{ij}$. In other words, U_i represents the number of survival (or censored) times which are *definitely* less than t_i (or t_i^+) minus the number of survival (or censored) times which are *definitely* greater than t_i (or t_i^+). Once we get the importance score for each individual, the next step is straightforward. We can just easily replace the original survival times $(x_{11}, \dots, x_{m1}, y_{11}, \dots, x_{n1},)$ with $(u_1, \dots, u_{m+n})$. We summarize the procedure in Algorithm 2.4.

```
Algorithm 2.4.  Input: $(x_1, \dots, x_m)$ and $(y_1, \dots, y_n)$.
  Pool $(x_{11}, \dots, x_{m1})$ and $(y_{11}, \dots, y_{n1})$ into a single group as $(t_1, \dots, t_{m+n})$.
```

Compute the importance score $u_i = \sum_{j=1}^{m+n} \widetilde{u}_{ij}$ *by* (2.8).
Replace $(x_{11}, \ldots, x_{m1})$ *and* $(y_{11}, \ldots, y_{n1})$ *with* $(u_1, \ldots, u_m)$ *and* $(u_{m+1}, \ldots, u_{m+n})$, *respectively.*
Compute test statistic $\mathrm{RE}^2_{m,n}$ *using Algorithm* 2.2 *with input* $(\widetilde{x}_1, \ldots, \widetilde{x}_m)$ *and* $(\widetilde{y}_1, \ldots, \widetilde{y}_n)$, *where* $\widetilde{x}_i = (u_i, x_{i2}, \ldots, x_{id})^\top, \widetilde{y}_j = (u_{m+j}, y_{j2}, \ldots, y_{jd})^\top$.
***Return**: test statistic* $\mathrm{RE}^2_{m,n}$.

In the first step, we utilize the generalized Wilcoxon pairwise comparisons proposed by Gehan [12] to calculate the relative rank of the survival term for each subject. It is important to note that there are other methods available for obtaining relative ranks, such as imputation-based approaches [39] or inverse probability of censoring weighting approach [9]. For a more comprehensive review of these methods, please refer to [40].

3 Simulation

In this section, we assess the finite sample performance of the global multivariate rank-based approach and compare it with two other rank-based approaches: O'Brien's rank-sum procedure [5] and Wittkowski's method [7]. In Wittkowski's method, the pairwise comparison is based on $\phi(r_{ij}) = 1(\sum_{k=1}^{d} r_{ij}^{(k)} > 0) - 1(\sum_{k=1}^{d} r_{ij}^{(k)} < 0)$, where a score of 1 is assigned if subject 1 has more favorable outcomes than subject 2.

3.1 Multiple Uncensored Outcomes

In this section, we conduct simulation studies to evaluate the performance of the test procedure on uncensored endpoints. We consider two scenarios in which we examine both continuous and discrete endpoints.

[label=]*scenario 1:* Suppose we collect $d = 8$ measurements to verify the efficacy of an intervention and the observed values for two arms follow

$$x_1, \ldots, x_m \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \Sigma), y_1, \ldots, y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(r\mu, \Sigma),$$

where $\mu = (1, 0.1, 0.2, 0.3, 0.1, 0.8, 0.1, 0)^\top$ represents the mean vector, $\sigma_{ij} = 1$ if $i = j$, and $\sigma_{ij} = \rho$ if $i \neq j$ represents the (i, j)th entry of the covariance matrix Σ. We consider $\rho = 0.3, 0.8$ and r from 1 to 3. In particular, $r = 1$ is used to examine the empirical size of the proposed test under H_0, and other values of r are used to check the empirical powers against alternatives. The target significance level is chosen as $\alpha = 0.05$. The results are summarized in Fig. 3. It can be observed that the type I error of all three methods is well controlled. However, the multivariate rank approach performs significantly better when $r > 0$, indicating its superiority under H_1. An interesting observation is that Wittkowski's method performs better than O'Brien's method when the correlation between each endpoint is stronger. *scenario 2:* In this scenario, we consider four correlated endpoints ($d = 4$) where three of them are continuous and one is discrete. Specifically, we let $x_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})^\top$, $y_j = (y_{j1}, y_{j2}, y_{j3}, y_{j4})^\top \in \mathbb{R}^4$, where

$$(x_{i1}, x_{i2}, x_{i3})^\top \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \Sigma), \quad (y_{j1}, y_{j2}, y_{j3})^\top \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu - r\nu, \Sigma),$$
$$x_{i4} \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p_x), \quad y_{j4} \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p_y),$$

where

$$p_x = \frac{\exp\{-3 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}\}}{1 + \exp\{-3 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})\}},$$
$$p_y = \frac{\exp\{-3 + \beta_1 y_{j1} + \beta_2 y_{j2} + \beta_3 y_{j3}\}}{1 + \exp\{-3 + \beta_1 y_{j1} + \beta_2 y_{j2} + \beta_3 y_{j3}\}},$$
$$\Sigma = \begin{bmatrix} 10^2 & 7 & 0.6 \\ 7 & 1 & 0.4 \\ 0.6 & 0.4 & 15^2 \end{bmatrix},$$

$(\beta_1, \beta_2, \beta_3)^\top = (0.1, 0.4, 0.1)^\top, \mu = (150, 6, 250)^\top, \nu = (10, 0.1, 10)^\top$.. For our analysis, we vary the treatment effect parameter r from 0 to 1. Specifically, we set $r = 0$ to examine the empirical size of the proposed test under the null hypothesis H_0. The results are summarized in Fig. 4. It can be observed that when including a discrete endpoint, the type I error of the multivariate rank approach and Wittkowski's methods can still be well controlled at the target significance level. However, O'Brien's method shows an inflated type I error. Additionally, the multivariate rank approach exhibits greater power compared to Wittkowski's method.

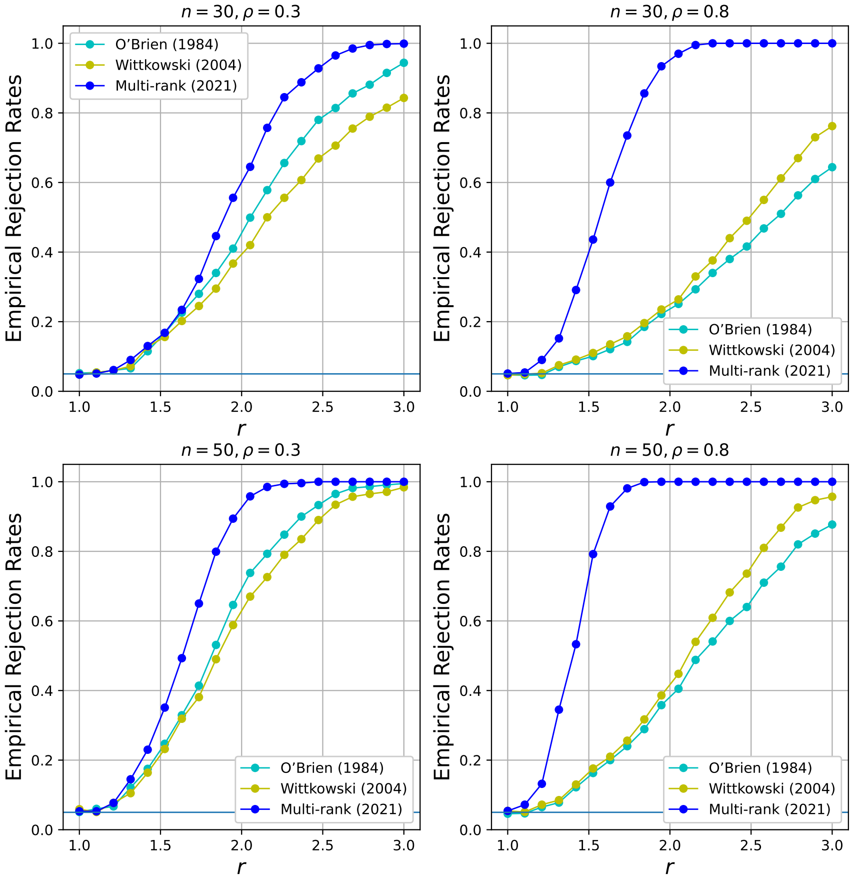

Fig. 3. Empirical rejection rates for scenario 1.

3.2 Time-to-Event Endpoint

In this section, we carry out simulation studies to assess the finite sample performance of the test procedure on right-censored survival data. Let $S_i^x(t|x_i) = \mathbb{P}(T_i^x > t|x_i), S_j^y(t|y_j) = \mathbb{P}(T_j^y > t|y_j)$ be the survical function, representing the probability of surviving beyond time t, where T_i^x, T_j^y are the survival times of subject x_i, y_j, respectively, and $x_i = (x_{i2}, \ldots, x_{id})^\top, y_j = (y_{j2}, \ldots, y_{jd})^\top$. We conciser the following Cox proportional hazards model

$$S_i^x(t|x_i) = \exp\left(-H_0(t)\exp(\psi(x_i))\right), \quad S_j^y(t|y_j) = \exp\left(-H_0(t)\exp(\psi(y_j))\right),$$

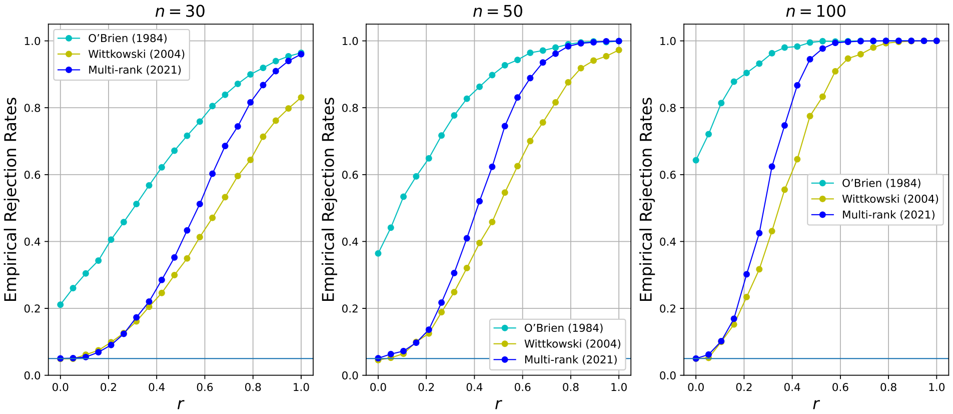

Fig. 4. Empirical rejection rates for scenario 2.

where $H_0(t)$ is the cumulative baseline hazard function and $\psi(x)$ is the covariates effect. We use the inverse probability method by [41] to generate T_i^x, T_j^y from the hazard function. Specifically, let U_i, U_j be uniformly distributed on $[0, 1]$, then the corresponding event time

$$T_i^x = (S_i^x)^{-1}(U_i|x_i) = H_0(t)^{-1}\left(-\frac{\log(U_i)}{\exp(\psi(x_i))}\right), T_j^y = (S_j^y)^{-1}(U_j|y_j) = H_0(t)^{-1}\left(-\frac{\log(U_j)}{\exp(\psi(y_j))}\right).$$

In this simulation, we consider the number of endpoints $d = 6$, where

$$x_i = (x_{i2}, \dots, x_{i6})^\top \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \Sigma), \quad y_j = (y_{j2}, \dots, y_{j6})^\top \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu - r\nu, \Sigma),$$

with $\mu = (3, 2, 2, 1, 1)^\top$, $\nu = (1, 0.1, 0, 0.1, 0.2)^\top$, $\sigma_{ij} = 1$ if $i = j$, and $\sigma_{ij} = \rho$ if $i \neq j$ representing the (i, j)th entry of the covariance matrix Σ. We consider $\rho = 0.3, 0.6$. In the context of the survival times T_i^x, T_j^y, we assume the baseline hazard function is constant, i.e. the survival times are exponentially distributed which are generated from

$$T_i^x = (S_i^x)^{-1}(U_i|x_i) = -\frac{\log(U_i)}{\lambda \exp(\beta^\top x_i)}, \quad T_j^y = (S_j^y)^{-1}(U_j|y_j) = -\frac{\log(U_j)}{\lambda \exp(\beta^\top y_j)},$$

with $\lambda = 0.1, \beta = (0.5, 0.2, 0.3, 0.3, 0.5)^\top$. The corresponding survival endpoint x_{i1} and y_{j1} are defined as $x_{i1} = \min\{T_i^x, C_i^x\}, y_{j1} = \min\{T_j^y, C_j^y\}$, where censoring times C_i^x, C_j^y are generated from a uniform distribution $U(0, 3)$ and $\delta_i^x = 1(T_i^x \leq C_i^x), \delta_j^y = 1(T_j^y \leq C_j^y)$ denote the censoring indicator. To extend O'Brien's and Wittkowski's methods to survival endpoints, similar to the description in Section 2.3, we first use Wilcoxon pairwise comparison to obtain the relative rank of the survival time for each subject. We then replace the survival endpoint with the corresponding relative rank. We applied Algorithm 2.4 to examine the empirical size and empirical power of three testing procedures, and the results are summarized in Fig. 5. It can be observed that when $r = 0$ (under H_0), the empirical rejection rates are all around 5%, indicating that the type I error is well controlled in all three testing procedures. However, as r exceeds a threshold, the testing procedure based on the multivariate rank shows significantly better performance compared to the other two methods. Furthermore, as the correlation between each endpoint becomes stronger, the difference in power becomes larger, and Wittkowski's method outperforms O'Brien's method when the correlation is stronger. These findings are consistent with the results observed in scenario 1. The results demonstrate the validity of Algorithm 2.4.

3.3 Sensitivity Analysis

In this section, we conduct a sensitivity analysis to examine the influence of the low-discrepancy sequence used in the construction of multivariate ranks, as described in Section 2. We compare four different methods: uniform number in $[0,1]^d$, Hammersley sequence [34], Halton sequences [35], and Sobol' sequences [36]. For the uniform number, each component is generated from a standard uniform distribution. We evaluate the empirical rejection rates using scenario 1 with $\rho = 0.8$. Fig. 6. presents the results, indicating that all methods effectively control the type I error. However, the low-discrepancy sequences (Hammersley, Halton, and Sobol') demonstrate higher power compared to the uniform numbers method. Importantly, the choice of low-discrepancy sequence does not significantly impact the performance. This finding is consistent with Fig. 2, which illustrates the more even distribution of points provided by low-discrepancy sequences compared to randomly generated points.

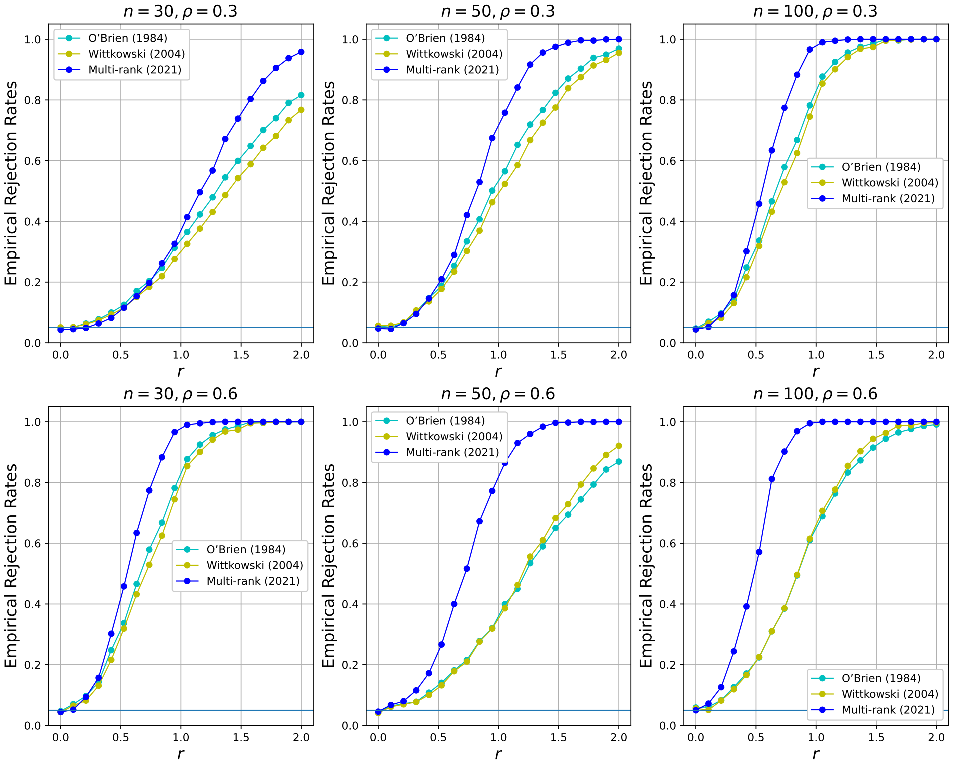

Fig. 5. Empirical rejection rates for the simulation of time-to-event endpoint in Section 3.2.

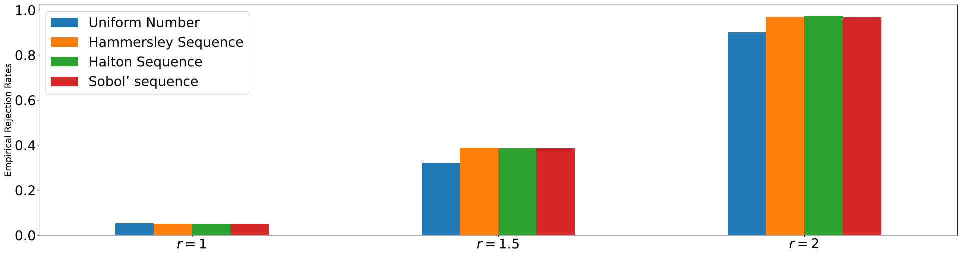

Fig. 6. Empirical rejection rates of scenario 1 for various value of r and $\rho = 0.8$ using different low-discrepancy methods.

4 Conclusion

In this study, we studied a global nonparametric testing procedure based on multivariate rank for the analysis of multiple endpoints in clinical trials. We compared the multivariate rank approach with other two existing rank-based methods, namely O'Brien's rank-sum procedure and Wittkowski's method. Through extensive simulations, we observed that the multivariate rank approach consistently outperformed the classical methods in terms of both type I error control and power. The use of multivariate rank allowed us to directly incorporate the relationships among multiple endpoints in the testing procedure, providing a more comprehensive and informative analysis. This approach exhibited robustness against various data distributions and censoring mechanisms commonly encountered in clinical trials. Additionally, we conducted sensitivity analyses to assess the impact of low-discrepancy sequences on the performance of the multivariate rank-based approach. The results showed that incorporating low-discrepancy sequences, such as Hammersley, Halton, and Sobol', further enhanced the power of the method without compromising its overall performance. In conclusion, our study highlights the utility of the multivariate rank-based approach for the analysis of multiple endpoints in clinical trials. By leveraging the relationships among endpoints, this method offers improved power and robustness compared to existing rank-based methods. Further research could explore the extension of these methods to handle additional complexities and real-world clinical trial datasets.

Acknowledgment

A brief acknowledgement section may be given after the conclusion section just before the references. The acknowledgments of people who provided assistance in manuscript preparation, funding for research, etc. should be listed in this section.

Competing Interests

Authors have declared that no competing interests exist.

References

[1] Bonferroni C. Teoria statistica delle classi e calcolo delle probabilita. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze. 1936;8:3-62.

[2] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological). 1995;57(1):289-300.

[3] Ristl R, Urach S, Rosenkranz G, Posch M. Methods for the analysis of multiple endpoints in small populations: A review. Journal of biopharmaceutical statistics. 2019;29(1):1-29.

[4] Deb N, Sen B. Multivariate rank-based distribution-free nonparametric testing using measure transportation. Journal of the American Statistical Association. Published online 2021:1-16.

[5] O'Brien PC. Procedures for comparing samples with multiple endpoints. Biometrics. Published online 1984:1079-1087.

[6] Finkelstein DM, Schoenfeld DA. Combining mortality and longitudinal measures in clinical trials. Statistics in medicine. 1999;18(11):1341-1354.

[7] Wittkowski KM, Lee E, Nussbaum R, Chamian FN, Krueger JG. Combining several ordinal measures in clinical studies. Statistics in medicine. 2004;23(10):1579-1592.

[8] Huang P, Woolson RF, O'Brien PC. A rank-based sample size method for multiple outcomes in clinical trials. Statistics in medicine. 2008;27(16):3084-3104.

[9] Buyse M. Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. Statistics in medicine. 2010;29(30):3245-3257.

[10] Berry JD, Miller R, Moore DH, et al. The Combined Assessment of Function and Survival (CAFS): a new endpoint for ALS clinical trials. Amyotrophic lateral sclerosis and frontotemporal degeneration. 2013;14(3):162-168.

[11] Ramchandani R, Schoenfeld DA, Finkelstein DM. Global rank tests for multiple, possibly censored, outcomes. Biometrics. 2016;72(3):926-935.

[12] Gehan EA. A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. Biometrika. 1965;52(1-2):203-224.

[13] Pocock SJ, Ariti CA, Collier TJ, Wang D. The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. European heart journal. 2012;33(2):176-182.

[14] Luo X, Tian H, Mohanty S, Tsai WY. An alternative approach to confidence interval estimation for the win ratio statistic. Biometrics. 2015;71(1):139-145.

[15] Dong G, Li D, Ballerstedt S, Vandemeulebroecke M. A generalized analytic solution to the win ratio to analyze a composite endpoint considering the clinical importance order among components. Pharmaceutical Statistics. 2016;15(5):430-437.

[16] Spearman C. The proof and measurement of association between two things. American Journal of Psychology. 1904;15(1):72-101.

[17] Cramér H. On the composition of elementary errors: First paper: Mathematical deductions. Scandinavian Actuarial Journal. 1928;1928(1):13-74.

[18] Smirnov NV. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. Bull Math Univ Moscou. 1939;2(2):3-14.

[19] Wald A, Wolfowitz J. On a test whether two samples are from the same population. The Annals of Mathematical Statistics. 1940;11(2):147-162.

[20] Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. The annals of mathematical statistics. Published online 1947:50-60.

[21] Hoeffding W. A non-parametric test of independence. The Collected Works of Wassily Hoeffding. Published online 1994:214-226.

[22] Bickel PJ. A distribution free version of the Smirnov two sample test in the p-variate case. The Annals of Mathematical Statistics. 1969;40(1):1-23.

[23] Friedman JH, Rafsky LC. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. The Annals of Statistics. Published online 1979:697-717.

[24] Maa JF, Pearl DK, Bartoszyński R. Reducing multidimensional two-sample data to one-dimensional interpoint comparisons. The annals of statistics. 1996;24(3):1069-1074.

[25] Gretton A, Borgwardt KM, Rasch MJ, Schölkopf B, Smola A. A kernel two-sample test. The Journal of Machine Learning Research. 2012;13(1):723-773.

[26] Chen H, Friedman JH. A new graph-based two-sample test for multivariate and object data. Journal of the American statistical association. 2017;112(517):397-409.

[27] Mukhopadhyay S, Wang K. A nonparametric approach to high-dimensional k-sample comparison problems. Biometrika. 2020;107(3):555-572.

[28] Liu F, Xu W, Lu J, Zhang G, Gretton A, Sutherland DJ. Learning deep kernels for non-parametric two-sample tests. In: International Conference on Machine Learning. PMLR; 2020:6316-6326.

[29] Liu F, Xu W, Lu J, Sutherland DJ. Meta two-sample testing: Learning kernels for testing with limited data. Advances in Neural Information Processing Systems. 2021;34:5848-5860.

[30] Oja H, Randles RH. Multivariate nonparametric tests. Statistical Science. 2004;19(4):598-605.

[31] Rosenbaum PR. An exact distribution-free test comparing two multivariate distributions based on adjacency. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2005;67(4):515-530.

[32] Biswas M, Mukhopadhyay M, Ghosh AK. A distribution-free two-sample run test applicable to high-dimensional data. Biometrika. 2014;101(4):913-926.

[33] McCann RJ. Existence and uniqueness of monotone measure-preserving maps. Duke Mathematical Journal. 1995;80(2).

[34] Hammersley JM. Monte Carlo methods for solving multivariable problems. Annals of the New York Academy of Sciences. 1960;86(3):844-874.

[35] Halton J, Smith GB. Radical inverse quasi-random point sequence, algorithm 247. Commun ACM. 1964;7(12):701.

[36] Sobol' IM. On the distribution of points in a cube and the approximate evaluation of integrals. Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki. 1967;7(4):784-802.

[37] Li K, Wang F, Yang L, Liu R. Deep feature screening: Feature selection for ultra high-dimensional data via deep neural networks. Neurocomputing. 2023;538:126186.

[38] Zhao S, Fu G. Distribution-free and model-free multivariate feature screening via multivariate rank distance correlation. Journal of Multivariate Analysis. 2022;192:105081.

[39] Efron B. The two sample problem with censored data. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. 1967; 4:831-853.

[40] Deltuvaite-Thomas V, Verbeeck J, Burzykowski T, et al. Generalized pairwise comparisons for censored data: An overview. Biometrical Journal. 2023;65(2):2100354.

[41] Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. Statistics in medicine. 2005;24(11):1713-1723.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
http://www.sdiarticle5.com/review-history/112095