*Biometrical Journal*

WILEY-VCH

Biometrical Journal

RESEARCH ARTICLE OPEN ACCESS

# Incompletely Observed Nonparametric Factorial Designs With Repeated Measurements: A Wild Bootstrap Approach

Lubna Amro[1] | Frank Konietschke[2,3] | Markus Pauly[1,4]

[1]Department of Statistics, TU Dortmund University, Dortmund, Germany | [2]Institute of Biometry and Clinical Epidemiology, Charitè—Universitätsmedizin Berlin, Berlin, Germany | [3]Berlin Institute of Health (BIH), Berlin, Germany | [4]UA Ruhr, Research Center Trustworthy Data Science and Security, Dortmund, Germany

**Correspondence:** Lubna Amro (lubna.amro@tu-dortmund.de)

**Received:** 23 May 2023 | **Revised:** 22 May 2024 | **Accepted:** 8 July 2024

**Funding:** This work is supported by the German Research Foundation (DFG) (No. PA 2409/3-2). Also, Frank Konietschke's work is supported by the German Research Foundation (DFG) (No. KO 4680/3-2) and Lubna Amro's work is supported by the project "From Prediction to Agile Interventions in the Social Sciences (FAIR)," which receives funding from the program "Profilbildung 2020," an initiative of the Ministry of Culture and Science of the State of Northrhine Westphalia.

**Keywords:** missing values | nonparametric hypotheses | ordered categorical data | rank tests | repeated measures | wild bootstrap

## ABSTRACT

In many life science experiments or medical studies, subjects are repeatedly observed and measurements are collected in factorial designs with multivariate data. The analysis of such multivariate data is typically based on multivariate analysis of variance (MANOVA) or mixed models, requiring complete data, and certain assumption on the underlying parametric distribution such as continuity or a specific covariance structure, for example, compound symmetry. However, these methods are usually not applicable when discrete data or even ordered categorical data are present. In such cases, nonparametric rank-based methods that do not require stringent distributional assumptions are the preferred choice. However, in the multivariate case, most rank-based approaches have only been developed for complete observations. It is the aim of this work to develop asymptotic correct procedures that are capable of handling missing values, allowing for singular covariance matrices and are applicable for ordinal or ordered categorical data. This is achieved by applying a wild bootstrap procedure in combination with quadratic form-type test statistics. Beyond proving their asymptotic correctness, extensive simulation studies validate their applicability for small samples. Finally, two real data examples are analyzed.

## 1 | Introduction

Factorial designs have a long history in several scientific fields such as biology, ecology, or medicine (Traux and Carkhuff 1965; Nesnow et al. 1998; Wildsmith et al. 2001; Biederman, Boutton, and Whisenant 2008; Lekberg et al. 2018). The reasons are simple: They are an efficient way to study main and interaction effects between different factors. Typically, such designs are inferred by parametric mean-based procedures such as linear mixed models, or multivariate analysis of variance (MANOVA). These procedures, however, rely on restrictive distributional assumptions, such as multivariate normality, or special dependencies (Lawley 1938; Bartlett 1939; Davis 2002; Johnson and Wichern 2007; Fitzmaurice, Laird, and Ware 2012).

However, assessing multivariate normality or a specific type of covariance matrix is difficult in practice (Micceri 1989; Qian and Huang 2005; Bourlier et al. 2008; Xu and Cui 2008), especially when the sample sizes are small. In particular, Erceg-Hurn and Mirosevich (2008) pointed out that "*Researchers relying on statis-*

*tical tests (e.g., Levene's test) to identify assumption violations may frequently fail to detect deviations from normality and homoscedasticity that are large enough to seriously affect the type-I error rate and power of classic parametric tests*" (p. 600). In fact, classical MANOVA procedures are usually nonrobust to deviations and may result in inaccurate decisions caused by possibly conservative or inflated type-I error rates (Arnau et al. 2012; Pesarin and Salmaso 2012; Brombin, Midena, and Salmaso 2013; Konietschke et al. 2015; Pauly, Ellenberger, and Brunner 2015; Friedrich, Brunner, and Pauly 2017; Friedrich and Pauly 2018). Moreover, if ordinal, or ordered categorical data are present, mean-based approaches are not applicable. A tempting alternative are nonparametric rank-based methods, which are applicable for nonnormal data, in particular discrete data or even ordered categorical data. Their key features are their robustness and their invariance under monotone transformations of the data. Consequently, Akritas and Arnold (1994), Akritas and Brunner (1997), Munzel and Brunner (2000), Brunner and Puri (2001), Akritas (2011), Brunner et al. (2017), and Friedrich, Konietschke, and Pauly (2017) proposed nonparametric ranking methods for all kinds of factorial designs. Thereby, hypotheses are no longer formulated in terms of means (which do not exist for ordinal data), distribution functions are used instead. However, all these methods are only applicable for completely observed factorial designs data and cannot be used to analyze multivariate data with missing values.

In contrast, there are only a few methods that are applicable in case of missing values, require no parametric assumptions, and also lead to valid inferences in case of arbitrary covariance structures, skewed distributions, or unbalanced experimental designs. Examples in the nonparametric framework for the matched pairs designs include the ranking methods proposed by Akritas, Kuha, and Osgood (2002), Konietschke et al. (2012), and Fong et al. (2018). A promising approach for factorial designs with repeated measurements is given by the proposals of Brunner, Munzel, and Puri (1999) and Domhof, Brunner, and Osgood (2002) who recommended two types of quadratic forms for testing nonparametric hypotheses in terms of distribution functions: rank-based Wald-type statistic (WTS) and ANOVA-type statistic (ATS). The Wald-type test is an asymptotically valid test, which usually needs large sample sizes to obtain accurate test decisions, see Brunner (2001) and Friedrich, Konietschke, and Pauly (2017) for the case of complete observations and our simulation study below. Apart from that, the ANOVA-type test is based on an approximation of its distribution with scaled $\chi^2$-distributions. As the latter does not coincide with the ANOVA-type test's limiting distribution under the null hypothesis $H_0$, the test is in general not asymptotically correct and also exhibits a more or less conservative behavior under small sample sizes, see Brunner (2001) for the case of complete observations. Another applicable technique is the nonparametric imputation method proposed by Gao (2007), which possesses a good type-I error control but a lower power behavior than the methods by Brunner, Munzel, and Puri (1999) and Domhof, Brunner, and Osgood (2002), see the simulation study of Gao (2007). Additionally, recent works by Rubarth, Pauly, and Konietschke (2022) and Rubarth et al. (2022) introduce promising inference methods for testing null hypotheses formulated in terms of relative effects instead of distribution functions in factorial designs with missing data, without relying on parametric assumptions. Rubarth, Pauly, and Konietschke (2022) focus on nonparametric techniques for repeated measures designs, while Rubarth et al. (2022) extend this research to factorial designs with clustered data.

The aims of the present paper are (i) to provide statistical tests for hypotheses formulated in distribution functions that are capable of treating missing values in factorial designs; (ii) work without parametric assumptions such as continuity of the distribution functions, or nonsingular covariance matrices; (iii) are asymptotically correct while (iv) showing a satisfactorily type-I error control and good power properties. To accomplish this, we propose three different quadratic-form–type test statistics and equip them with a nonparametric Wild bootstrap procedure for calculating critical values to enhance their small sample performance. Here, we throughout assume missing completely at random (MCAR) mechanism, when constructing the test statistics and developing their related theories. However, in the simulation studies, we investigate the effects of some missing at random (MAR) scenarios on the test statistics performance under small sample sizes, see the Supporting Information for the explicit definition of the missing mechanisms.

The paper is organized as follows: First we introduce the statistical model and the hypotheses of interest in the next section. In Section 3, we introduce the test statistics and analyze their asymptotic behavior. In Section 4, the proposed wild bootstrap technique is explained. Section 5 displays the results from our extensive simulation study and two real data examples from a fluvoxamine study and a skin disorder clinical trial are analyzed in Section 6. All proofs as well as additional simulation results can be found in the Supporting Information.

To facilitate the presentation, we introduce the following notations: Let $\boldsymbol{I}_d$ denote the $d$-dimensional unit matrix, $\boldsymbol{J}_d$ denote the $d \times d$ matrix of $\mathbf{1}$s that is, $\boldsymbol{J}_d = \mathbf{1}_d \mathbf{1}_d^T$, where $\mathbf{1}_d = (1, \dots, 1)_{d\times 1}^T$ denotes the $d$-dimensional column vector and $\boldsymbol{P}_d = \boldsymbol{I}_d - \frac{1}{d}\boldsymbol{J}_d$ is the so-called $d$-dimensional centering matrix. Finally, by $\boldsymbol{A} \otimes \boldsymbol{B}$ we denote the Kronecker product of the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$.

## 2 | Statistical Model and Nonparametric Hypotheses

We consider a nonparametric repeated measures model with $a$ independent and potentially unbalanced treatment groups and $d$ different time points given by independent random vectors

$$\boldsymbol{X}_{ik} = (X_{i1k}, \dots, X_{idk})^T, \quad i = 1, \dots, a, \quad k = 1, \dots, n_i, \tag{2.1}$$

where $X_{ijk} \sim F_{ij}(x) = \frac{1}{2}[F_{ij}^+(x) + F_{ij}^-(x)], i = 1, \dots, a, j = 1, \dots, d, k = 1, \dots, n_i$. Here, $F_{ij}^+(x) = P(X_{ijk} \le x)$ is the right continuous version and $F_{ij}^-(x) = P(X_{ijk} < x)$ is the left continuous version of the distribution function. Using the normalized version $F_{ij}(x)$ is useful for handling ties and including continuous as well as discontinuous distribution functions (Brunner, Bathke, and Konietschke 2018). To include the case of missing values, we follow the notation of Brunner, Munzel, and Puri (1999) and let

$$\lambda_{ijk} = \begin{cases} 1, & \text{if } X_{ijk} \text{ is observed} \\ 0, & \text{if } X_{ijk} \text{ is nonobserved} \end{cases}$$
$$i = 1, \dots, a, j = 1, \dots, d, k = 1, \dots, n_i. \tag{2.2}$$

Moreover, let $n = \sum_{i=1}^{a} n_i$ denote the total number of subjects and let

$N = \sum_{i=1}^{a} \sum_{j=1}^{d} \sum_{k=1}^{n_i} \lambda_{ijk}$ denote the total number of observations.

To formulate the null hypothesis in this nonparametric setup, let $\boldsymbol{F} = (F_{11}, \dots, F_{ad})^T$ denote the vector of the distribution functions $F_{ij}, i = 1, \dots, a, j = 1, \dots, d.$ and $\boldsymbol{C}$ denote a contrast matrix, that is, $\boldsymbol{C}\mathbf{1} = \mathbf{0}$ where $\mathbf{1} = (1, \dots, 1)^T$ and $\mathbf{0} = (0, \dots, 0)^T$. Then, the null hypotheses are formulated by $H_0 : \{\boldsymbol{CF} = \mathbf{0}\}$. This framework covers different factorial repeated measures designs. For example, the hypothesis of no treatment group effect, that is, $H_0^G : \{\bar{F}_{1\cdot} = \cdots = \bar{F}_{a\cdot}\}$, is equivalently written in matrix notation as $H_0^G : \{(\boldsymbol{P}_a \otimes \frac{1}{d}\mathbf{1}_d^T)\boldsymbol{F} = \mathbf{0}\}$. Similarly, the hypothesis of no time effect, that is, $H_0^T : \{\bar{F}_{\cdot 1} = \cdots = \bar{F}_{\cdot d}\}$ is equivalently written as $H_0^T : \{(\frac{1}{a}\mathbf{1}_a^T \otimes \boldsymbol{P}_d)\boldsymbol{F} = \mathbf{0}\}$, and the hypothesis of no interaction effect between treatment and time is written as $H_0^{GT} : \{(\boldsymbol{P}_a \otimes \boldsymbol{P}_d)\boldsymbol{F} = \mathbf{0}\}$.

We note that more complex factorial structures on the repeated measures (e.g., in case of different interventions over time as in Sattler and Pauly 2018 or the groups (in case of two or more grouping factors) are also covered by our approach by simply splitting up the indices $i$ (for a factorial group structure) or $j$ (for a factorial time structure). For ease of presentation, we will focus on the above hypotheses.

To entail a parameter for describing differences between distributions, Brunner, Munzel, and Puri (1999) and Domhof, Brunner, and Osgood (2002) considered the relative marginal effects

$$p_{ij} = \int H(x) dF_{ij}(x), \tag{2.3}$$

where $H(x) = N^{-1} \sum_{i=1}^{a} \sum_{j=1}^{d} \sum_{k=1}^{n_i} \lambda_{ijk} F_{ij}(x)$ is the weighted average of all distribution functions in the experiment. Estimators thereof are given by plugging-in the empirical versions of $F_{ij}(x)$ and $H(x)$

$$\hat{F}_{ij}(x) = \frac{1}{\lambda_{ij\cdot}} \sum_{k=1}^{n_i} \lambda_{ijk} \hat{F}_{ijk}(x) \tag{2.4}$$

$$= \frac{1}{\lambda_{ij\cdot}} \sum_{k=1}^{n_i} \lambda_{ijk} c(x - X_{ijk}), \quad \lambda_{ij\cdot} = \sum_{k=1}^{n_i} \lambda_{ijk}, \tag{2.5}$$

$$\hat{H}(x) = \frac{1}{N} \sum_{i=1}^{a} \sum_{j=1}^{d} \sum_{k=1}^{n_i} c(x - X_{ijk}), \tag{2.6}$$

where $c(u)$ is the normalized version of the counting function, that is, $c(u) = 0, 1/2$ or $1$ according as $u < 0, u = 0$ or $u > 0$ and $c(x - X_{ijk})$ is assumed to equal 0 if the observation $X_{ijk}$ is missing. Note that $\hat{F}_{ij}(x) = 0$ in case of $\lambda_{ij\cdot} = 0$. Thus, the relative marginal effects $p_{ij}$ are estimated by

$$\hat{p}_{ij} = \int \hat{H} d\hat{F}_{ij} = \frac{1}{\lambda_{ij\cdot}} \sum_{k=1}^{n_i} \lambda_{ijk} \hat{H}(X_{ijk}) \tag{2.7}$$

$$= \frac{1}{\lambda_{ij\cdot}} \sum_{k=1}^{n_i} \lambda_{ijk} \frac{1}{N} \left( R_{ijk} - \frac{1}{2} \right), \tag{2.8}$$

where $R_{ijk}$ is the mid-rank of $X_{ijk}$ among all $N$ dependent and independent observations. For convenience, the relative marginal effect estimators are collected in the vector $\hat{\boldsymbol{p}} = (\hat{p}_{11}, \dots, \hat{p}_{ad})^T$.

In the sequel, we derive asymptotic theory under the following sample size assumption and missing values:

**Assumption 1.** $\frac{\lambda_{ij\cdot}}{n} \to \kappa_i \in [0, 1] \quad i = 1, \dots, a, \quad j = 1, \dots, d,$ as $n_0 \equiv min\{\lambda_{ij\cdot}\} \to \infty$.

This ensures that there are "enough" subjects without missing values in each group.

## 2.1 | Asymptotic Distribution

Brunner, Munzel, and Puri (1999) derived the asymptotic distribution of $\sqrt{n}\boldsymbol{C}\hat{\boldsymbol{p}}$ under the null hypothesis $H_0 : \boldsymbol{CF} = \mathbf{0}$. Setting $\hat{\boldsymbol{Y}}_{ijk} = \hat{H}(X_{ijk})$ and $\boldsymbol{Y}_{ijk} = H(X_{ijk})$, they proved that under $H_0$, $\sqrt{n}\boldsymbol{C}\hat{\boldsymbol{p}}$ and $\sqrt{n}\boldsymbol{C}\bar{\boldsymbol{Y}}_{\cdot}$ are asymptotically equivalent. Accordingly, they showed that $\sqrt{n}\boldsymbol{C}\hat{\boldsymbol{p}}$ follows under $H_0$, asymptotically, a multivariate normal distribution with expectation $\mathbf{0}$ and covariance matrix $\boldsymbol{C}\boldsymbol{V}_n\boldsymbol{C}^T$. Here

$$\boldsymbol{V}_n = \bigoplus_{i=1}^{d} \frac{n}{n_i} \boldsymbol{V}_i = \bigoplus_{i=1}^{d} \frac{n}{n_i^2} \sum_{k=1}^{n_i} \boldsymbol{\Lambda}_{ik} \boldsymbol{V}_{ik} \boldsymbol{\Lambda}_{ik}, \tag{2.9}$$

where $\boldsymbol{\Lambda}_{ik} = n_i diag\{\frac{\lambda_{i1k}}{\lambda_{i1\cdot}}, \dots, \frac{\lambda_{idk}}{\lambda_{id\cdot}}\}$ and $\boldsymbol{V}_{ik} = Cov(\boldsymbol{Y}_{ik})$ denotes the covariance matrix of the random vectors $\boldsymbol{Y}_{ik} = (H(X_{i1k}), \dots, H(X_{idk}))^T$.

Brunner, Munzel, and Puri (1999) propose to estimate the covariance matrix $\boldsymbol{V}_n$ by

$$\hat{\boldsymbol{V}}_n = \bigoplus_{i=1}^{r} \frac{n}{n_i} \hat{\boldsymbol{V}}_i, \tag{2.10}$$

where $\hat{\boldsymbol{V}}_i = [\hat{v}_i(j, j')]$ with diagonal and off-diagonal elements $\hat{v}_i(j, j)$ and $\hat{v}_i(j, j')$, respectively, defined as

$$\hat{v}_i(j, j) = \frac{n_i \sum_{k=1}^{n_i} \lambda_{ijk} [R_{ijk} - \bar{R}_{ij\cdot}]^2}{(N^2)\lambda_{ij\cdot}(\lambda_{ij\cdot} - 1)}, \tag{2.11}$$

$$\hat{v}_i(j, j') = \frac{n_i \sum_{k=1}^{n_i} \lambda_{ijk}\lambda_{ij'k} [(R_{ijk} - \bar{R}_{ij\cdot})(R_{ij'k} - \bar{R}_{ij'\cdot})]}{(N^2)((\lambda_{ij\cdot} - 1)(\lambda_{ij'\cdot} - 1) + \Delta_{i,jj'} - 1)}. \tag{2.12}$$

Here, $\bar{R}_{ij\cdot} = \frac{1}{\lambda_{ij\cdot}} \sum_{k=1}^{n_i} \lambda_{ijk}(R_{ijk})$ and $\Delta_{i,jj'} = \sum_{k=1}^{n_i} \lambda_{ijk}\lambda_{ij'k}$. Under Assumption 1, Brunner, Munzel, and Puri (1999) have shown that $\hat{\boldsymbol{V}}_i$ is a consistent estimator of $\boldsymbol{V}_i$.

## 3 | Statistics and Asymptotics

In this section, we propose three different quadratic forms for testing the null hypothesis: a WTS, an ATS as suggested in Brunner, Munzel, and Puri (1999) and Domhof, Brunner, and Osgood (2002) as well as a modified ANOVA-type statistic (MATS). For mean-based analyses, the latter was proposed by Friedrich and Pauly (2018).

The rank version of the WTS is defined as

$$T_W = n\hat{\boldsymbol{p}}^T\boldsymbol{C}^T[\boldsymbol{C}\hat{\boldsymbol{V}}_n\boldsymbol{C}^T]^+\boldsymbol{C}\hat{\boldsymbol{p}}, \tag{3.1}$$

where $[\boldsymbol{B}]^+$ denotes the Moore-penrose inverse of a matrix $\boldsymbol{B}$. Its asymptotic null distribution is summarized below.

**Theorem 3.1.** *Under Assumption (1) and $\boldsymbol{V}_n \to \boldsymbol{V} > 0$ as $n_0 \to \infty$, the statistic $T_W$ has under the null hypothesis $H_0 : \boldsymbol{CF} = \boldsymbol{0}$, asymptotically, as $n_0 \to \infty$, a central $\chi^2_f$-distribution with $f = rank(\boldsymbol{C})$ degrees of freedom.*

WTSs of similar form are used in many different situations, for example, in heteroscedastic mean-based analyses (Krishnamoorthy and Lu 2010; Xu et al. 2013; Konietschke et al. 2015; Friedrich and Pauly 2018; Amro, Pauly, and Ramosaj 2019) and even more complex regression models in survival analyses (Martinussen and Scheike 2007; Dobler, Pauly, and Scheike 2019). However, the convergence of the WTS to its limiting $\chi^2$-distribution is usually slow and large sample sizes are required to obtain adequate results (Vallejo, Fernández, and Livacic-Rojas 2010; Konietschke et al. 2015; Pauly, Brunner, and Konietschke 2015; Smaga 2017). Thus, Brunner, Munzel, and Puri (1999) and Domhof, Brunner, and Osgood (2002) proposed an alternative quadratic form by deleting the variance $\hat{\boldsymbol{V}}_n$ involved in the computation of the WTS, resulting in the ATS defined as

$$T_A = \frac{1}{tr(\boldsymbol{T}\hat{\boldsymbol{V}}_n)} n\hat{\boldsymbol{p}}^T\boldsymbol{T}\hat{\boldsymbol{p}}, \tag{3.2}$$

where $\boldsymbol{T} = \boldsymbol{C}^T[\boldsymbol{C}\boldsymbol{C}^T]^+\boldsymbol{C}$ is a projection matrix. Note that $H_0 : \boldsymbol{TF} = \boldsymbol{0} \Leftrightarrow \boldsymbol{CF} = \boldsymbol{0}$ because $\boldsymbol{C}^T[\boldsymbol{C}\boldsymbol{C}^T]^+$ is a generalized inverse of $\boldsymbol{C}$.

It is also worthy to note that different to $T_W$, the asymptotic distribution of $T_A$ can also be derived if $\boldsymbol{V}$ is singular.

**Theorem 3.2.** *Under Assumption (1) and under the null hypothesis $H_0 : \boldsymbol{CF} = \boldsymbol{0}$, the test statistic $T_A$ has asymptotically, as $n_0 \to \infty$, the same distribution as the random variable*

$$\boldsymbol{A} = \sum_{i=1}^{a}\sum_{j=1}^{d} \zeta_{ij} B_{ij} / tr(\boldsymbol{T}\boldsymbol{V}_n), \tag{3.3}$$

*where $B_{ij} \overset{i.i.d}{\sim} \chi^2_1$ and the weights $\zeta_{ij}$ are the eigenvalues of $\boldsymbol{T}\boldsymbol{V}_n$.*

The limiting distribution of $\boldsymbol{A}$ is approximated by a scaled $g\chi^2_f$ distribution, where $g$ is a constant such that the first two moments coincide. Brunner, Munzel, and Puri (1999) proposed a Box (1954)–type approximation such that the first two moments of $T_A$ and $g\chi^2_f$ approximately coincide where $f$ can be estimated by $\hat{f} = \frac{(tr(\boldsymbol{T}\hat{\boldsymbol{V}}_n))^2}{tr(\boldsymbol{T}\hat{\boldsymbol{V}}_n\boldsymbol{T}\hat{\boldsymbol{V}}_n)}$. Thus, the distribution of $T_A$ is approximated by a central $F(\hat{f}, \infty)$-distribution. The corresponding ANOVA-type test $\phi_A = \mathbb{1}\{T_A > F_\alpha(\hat{f}, \infty)\}$, where $F_\alpha(\hat{f}, \infty)$ denotes the $(1-\alpha)$-quantile of the $F(\hat{f}, \infty)$-distribution.

Despite its advantage of being applicable in case of singular covariance matrices, the ATS has the drawback of being an approximative test and thus, even its asymptotic exactness cannot be guaranteed.

Another possible test statistic is the modified version of the ATS (MATS) that was developed by Friedrich and Pauly (2018) for mean-based MANOVA models. Here, we transfer it to the ranked-based set-up, where it is given by

$$T_M = n\hat{\boldsymbol{p}}^T\boldsymbol{C}^T[\boldsymbol{C}\hat{\boldsymbol{D}}_n\boldsymbol{C}^T]^+\boldsymbol{C}\hat{\boldsymbol{p}}, \tag{3.4}$$

where $\hat{\boldsymbol{D}}_n = diag(\frac{n}{n_i}\hat{v}_i(j,j)), i = 1,\dots,a, j = 1,\dots,d$. In this way, it is a compromise between the WTS and the ATS as it only uses the diagonal entries of $\hat{\boldsymbol{V}}_n$ for the multivariate studentization. In fact, the nonsingularity assumption of $\boldsymbol{V}$ is not needed to derive its asymptotics. It is replaced by the weaker requirement that $\mathbf{D} = diag(\frac{n}{n_i}v_i(j,j)) > 0, i = 1,\dots,a, j = 1,\dots,d$, which is fulfilled when all the diagonal elements $v_i(j,j)$ of $\boldsymbol{V}_i$ are positive. This is the same as $Var(H(X_{ij1})) > 0$ for all $i \in \{1,\dots,a\}, j \in \{1,\dots,d\}$. It is supposed to be met in a wide range of applicable settings, except only cases where, for example, at least one time point of any of the treatment groups is a discrete variable with very few distinct values.

**Theorem 3.3.** *Under Assumption (1) and assuming that $v_i(j,j) > 0$ for all $i \in \{1,\dots,a\}, j \in \{1,\dots,d\}$, the test statistic $T_M$, has under the null hypothesis $H_0 : \boldsymbol{CF} = \boldsymbol{0}$, asymptotically, as $n_0 \to \infty$, the same distribution as the random variable*

$$\boldsymbol{M} = \sum_{i=1}^{a}\sum_{j=1}^{d} \tilde{\zeta}_{ij}\tilde{B}_{ij},$$

*where $\tilde{B}_{ij} \overset{i.i.d}{\sim} \chi^2_1$ and the weights $\tilde{\zeta}_{ij}$ are the eigenvalues of $\boldsymbol{V}_n^{1/2}\boldsymbol{C}^T[\boldsymbol{C}\boldsymbol{D}\boldsymbol{C}^T]^+\boldsymbol{C}\boldsymbol{V}_n^{1/2}$ and $\mathbf{D} = diag(\frac{n}{n_i}v_i(j,j))$.*

Since, the limit distributions of both, the ATS and the MATS, depend on unknown weights $\zeta_i$ and $\tilde{\zeta}_i$, we cannot directly calculate critical values. In addition, the $\chi^2_f$-approximation to $T_W$ is rather slow. To this end, we develop asymptotically correct testing procedures based on bootstrap versions of $T_W$, $T_A$, and $T_M$ in the subsequent section.

## 4 | Wild Bootstrap Approach

We consider a wild bootstrap approach to derive new asymptotically valid testing procedures with good finite sample properties. To this end, let $\boldsymbol{Z}_{ik} = (\boldsymbol{R}_{ik} - \bar{\boldsymbol{R}}_{i.})$ denote the centered rank vectors, where $\boldsymbol{R}_{ik} = (R_{i1k},\dots,R_{idk})^T$, $\bar{\boldsymbol{R}}_{i.} = (\bar{R}_{i1.},\dots,\bar{R}_{id.})^T$, and $\bar{R}_{ij.} = \frac{1}{\lambda_{ij.}}\sum_{k=1}^{n_i}\lambda_{ijk}(R_{ijk})$. Moreover, let $W_{ik}$ denote independent and identically distributed random weights with $E(W_{ik}) = 0$ and $Var(W_{ik}) = 1$. Although, there are different possible choices for these random weights (Mammen 1993; Davidson and Flachaire 2008), some particular choices have become popular. Following the investigations in Friedrich, Konietschke, and Pauly (2017) for the corresponding complete case scenario, and our simulation results for other common wild bootstrap weights in Section 5.4.1 below, we use Rademacher random variables, which are defined by $P(W_{ik} = -1) = P(W_{ik} = 1) = 1/2$. Then, a wild

bootstrap sample is defined as

$$\boldsymbol{Z}_{ik}^{*} = W_{ik}.\boldsymbol{Z}_{ik}, \quad i = 1, \ldots, a, \quad k = 1, \ldots, n_i. \tag{4.1}$$

In other words, $\boldsymbol{Z}_{ik}^{*}$ is a symmetrization of the rank vector $\boldsymbol{Z}_{ik}$. Now, we can define the bootstrap version of the relative effect estimator $\hat{p}_{ij}$

$$\hat{p}_{ij}^{*} = \frac{1}{\lambda_{ij.}} \sum_{k=1}^{n_i} \lambda_{ijk} \frac{1}{N}(Z_{ijk}^{*}) = \frac{1}{\lambda_{ij.}} \sum_{k=1}^{n_i} \lambda_{ijk} \frac{1}{N}(W_{ik}(R_{ijk} - \bar{R}_{ij.})). \tag{4.2}$$

We pool them into the vector $\hat{\boldsymbol{p}}^{*} = (\hat{p}_{11}^{*}, \ldots, \hat{p}_{ad}^{*})^{T}$ to obtain the bootstrap counterpart of $\hat{\boldsymbol{p}}$.

In the same way, the bootstrap covariance matrix estimator $\hat{\boldsymbol{V}}_i^{*} = [\hat{v}_i^{*}(j, j')]$ with diagonal and off-diagonal elements $\hat{v}_i^{*}(j, j)$ and $\hat{v}_i^{*}(j, j')$, respectively, is given by

$$\hat{v}_i^{*}(j, j) = \frac{n_i \sum_{k=1}^{n_i} \lambda_{ijk}[Z_{ijk}^{*} - \bar{Z}_{ij.}^{*}]^2}{(N^2)\lambda_{ij.}(\lambda_{ij.} - 1)}, \tag{4.3}$$

$$\hat{v}_i^{*}(j, j') = \frac{n_i \sum_{k=1}^{n_i} \lambda_{ijk}\lambda_{ij'k}[(Z_{ijk}^{*} - \bar{Z}_{ij.}^{*})(Z_{ij'k}^{*} - \bar{Z}_{ij'.}^{*})]}{(N^2)((\lambda_{ij.} - 1)(\lambda_{ij'.} - 1) + \Delta_{i,jj'} - 1)}. \tag{4.4}$$

From this, the bootstrapped versions of the quadratic forms, that is, the WTS $T_W^{*}$, the ATS $T_A^{*}$, and the MATS $T_M^{*}$ are computed as

$$T_W^{*} = n\hat{\boldsymbol{p}}^{*T}\boldsymbol{C}^{T}[\boldsymbol{C}\hat{\boldsymbol{V}}_n^{*}\boldsymbol{C}^{T}]^{+}\boldsymbol{C}\hat{\boldsymbol{p}}^{*}, \tag{4.5}$$

$$T_A^{*} = \frac{1}{tr(\boldsymbol{T}\hat{\boldsymbol{V}}_n^{*})} n\hat{\boldsymbol{p}}^{*T}\boldsymbol{T}\hat{\boldsymbol{p}}^{*}, \tag{4.6}$$

$$T_M^{*} = n\hat{\boldsymbol{p}}^{*T}\boldsymbol{C}^{T}[\boldsymbol{C}\hat{\boldsymbol{D}}_n^{*}\boldsymbol{C}^{T}]^{+}\boldsymbol{C}\hat{\boldsymbol{p}}^{*}, \tag{4.7}$$

where $\hat{\boldsymbol{V}}_n^{*} = \bigoplus_{i=1}^{a} \frac{n}{n_i}\hat{\boldsymbol{V}}_i^{*}$ and $\hat{\boldsymbol{D}}_n^{*} = diag(\frac{n}{n_i}\hat{v}_i^{*}(j, j)), i = 1, \ldots, a$, $j = 1, \ldots, d$. To get an asymptotically valid bootstrap test, we have to assure that the conditional distribution of the Wald, ANOVA-type, and MATS-type bootstrap statistics $T_W^{*}$, $T_A^{*}$, and $T_M^{*}$ approximate the null distribution of $T_W$, $T_A$, and $T_M$, respectively.

**Theorem 4.1.** *Under Assumption (1), the following results hold:*

*(1) For $i = 1, \ldots, a$, the conditional distribution of $\sqrt{n}\hat{\boldsymbol{p}}_i^{*}$, given the data $\boldsymbol{X}$, converges weakly to the multivariate $N(0, \kappa_i^{-1}\boldsymbol{V}_i)$-distribution in probability.*

*(2) The conditional distribution of $\sqrt{n}\hat{\boldsymbol{p}}^{*}$, given the data $\boldsymbol{X}$, converges weakly to the multivariate $N(0, \bigoplus_{i=1}^{a} \kappa_i^{-1}\boldsymbol{V}_i)$-distribution in probability.*

Thus, the distributions of $\sqrt{n}\boldsymbol{C}\hat{\boldsymbol{p}}^{*}$ and $\sqrt{n}\boldsymbol{C}(\hat{\boldsymbol{p}} - \boldsymbol{p})$ asymptotically coincide under the null hypothesis $H_0$.

**Theorem 4.2.** *Under Assumption (1), for any choice $[-] \in \{A, M, W\}$, the conditional distribution of $T_{[-]}^{*}$ converges weakly to the null distribution of $T_{[-]}$ in probability for any choice of $\boldsymbol{p} \in \mathbb{R}^{ad}$. In particular, we have that*

$$\sup_{x \in \mathbb{R}} |P_{\boldsymbol{p}}(T_{[-]}^{*} \leq x | \boldsymbol{X}) - P_{H_0}(T_{[-]} \leq x)| \xrightarrow{p} 0$$

*holds, provided that $\boldsymbol{V} > 0$ or $v_i(j, j) > 0$ for all $i \in \{1, \ldots, a\}$, $j \in \{1, \ldots, d\}$ is fulfilled in case of $T_W$ or $T_M$, respectively.*

Therefore, the corresponding wild bootstrap tests are given by $\phi_W^{*} = \mathbb{1}\{T_W > c_w^{*}\}$, $\phi_A^{*} = \mathbb{1}\{T_A > c_A^{*}\}$, and $\phi_M^{*} = \mathbb{1}\{T_M > c_M^{*}\}$, where $c_w^{*}$, $c_A^{*}$, and $c_M^{*}$ denote the $(1 - \alpha)$ quantiles of the conditional bootstrap distributions of $T_W^{*}$, $T_A^{*}$, and $T_M^{*}$ given the data, respectively.

Theorem 4.2 implies that the wild bootstrap tests are of asymptotic level $\alpha$ under the null hypothesis and consistent for any fixed alternative $H_1 : \boldsymbol{CF} \neq \boldsymbol{0}$, that is, they have asymptotically power 1. In addition, it follows from Jansen et al. (2003) that they have the same local power under contiguous alternatives as their original tests.

## 5 | Monte Carlo Simulations

The above results are valid for large sample sizes. To analyze the finite sample behavior of the asymptotic quadratic tests and their wild bootstrap counterparts described in Sections 3 and 4, we conduct extensive simulations. As an assessment criteria, all procedures were studied with respect to their

(i) type-I error rate control at level $\alpha = 5\%$ and their

(ii) power to detect deviations from the null hypothesis.

All simulations were operated by means of the R computing environment, version 3.2.3 (R Core Team 2013) and each setting was based on 10,000 simulation runs and $B = 999$ bootstrap runs. The algorithm for the computation of the $p$-value of the wild bootstrap tests in any of the test statistics $T \in \{T_A, T_W, T_M\}$ is as follows:

(1) For the given incomplete multifactor data, calculate the observed test statistic, say $T$.

(2) Compute the rank values $\boldsymbol{Z}_{ik}$.

(3) Using i.i.d Rademacher weights $W_{i1}, \ldots, W_{in_i}$, generate bootstrapped rank values $\boldsymbol{Z}_{ik}^{*} = W_{ik}.\boldsymbol{Z}_{ik}$.

(4) Calculate the value of the test statistic for the bootstrapped sample $T^{*}$.

(5) Repeat the Steps 3 and 4 independently $B = 999$ times and collect the observed test statistic values in $T_b^{*}, b = 1, \ldots, B$.

(6) Finally, estimate the wild bootstrap $p$-value as $p\text{-value} = \frac{\sum_{b=1}^{B} I(T_b^{*} >= T)}{B}$.

We considered a two-way layout design with $a = 2$ independent groups and two different time points $d \in \{4, 8\}$ underlying discrete and continuous distributions. As in Konietschke et al. (2015) and Bathke et al. (2018) for the case of complete observations, we investigated balanced situations with sample size vectors $(n_1, n_2) \in \{(5, 5), (10, 10), (20, 20)\}$, and an unbalanced situation with sample size vector $(n_1, n_2) = (10, 20)$. Since the empirical

type-I error rates of the tests in sample sizes (10, 20) and (20, 10) are very similar, we omit (20, 10). In particular, we studied three different kinds of hypotheses: the hypothesis of no group effect $H_0^G : (\boldsymbol{P}_a \otimes \frac{1}{d}\mathbf{1}_d^T)$, no time effect $H_0^T : (\frac{1}{a}\mathbf{1}_a^T \otimes \boldsymbol{P}_d)$, as well as no interaction effect $H_0^{GT} : (\boldsymbol{P}_a \otimes \boldsymbol{P}_d)$.

For each scenario, we generated missingness within MCAR as well as MAR frameworks as described below: For the MCAR mechanism, we simulated $\boldsymbol{X}_{ik} = (\delta_{i1k}X_{i1k}, \ldots, \delta_{idk}X_{idk}), \quad k = 1, \ldots, n_i$, for independent Bernoulli distributed $\delta_{i1k} \sim B(r)$ and a zero entry was interpreted as a missing observation. The missing probability was chosen from $r \in \{10\%, 30\%\}$.

For the *MAR mechanism*, we considered several pairs of features $\{X_{obs}, X_{miss}\}$. For each pair, there is a determining feature $X_{obs}$ that determines the missing pattern of its corresponding $X_{miss}$ (Santos et al. 2019). Thus, for $d = 4$, we define the pairs $\{X_{i1k}, X_{i2k}\}$ and $\{X_{i3k}, X_{i4k}\}$. Whereas, for $d = 8$, we define the following pairs $\{X_{i1k}, X_{i2k}\}, \{X_{i1k}, X_{i3k}\}, \{X_{i6k}, X_{i7k}\}$, and $\{X_{i6k}, X_{i8k}\}$. Two different MAR scenarios are investigated—MAR1 and MAR2. In MAR1, for each pair, we divided $\mathbf{X}$ into three groups based on their $X_{obs}$ values: The first group is given by $\{\boldsymbol{X}_{ik} : X_{obs} \in (-\infty, -2\hat{\sigma}_1), k = 1, .., n_i\}$, the second by $\{\boldsymbol{X}_{ik} : X_{obs} \in (-2\hat{\sigma}_1, 2\hat{\sigma}_1), k = 1, .., n_i\}$, and the last group by $\{\boldsymbol{X}_{ik} : X_{obs} \in (2\hat{\sigma}_1, \infty), k = 1, .., n_i\}$, where $\hat{\sigma}_1$ is the estimated standard deviation of $X_{obs}$. Then, we randomly inserted missing values on $X_{miss}$ based on the following missing percentages: 15% for group 1 and three and 30% for the second group.

For generating MAR2 scenario, we considered the median of each $X_{obs}$ to define the missing pattern of $X_{miss}$ (Zhu, He, and Liatsis 2012; Pan et al. 2015). For each pair from above, two groups were defined: The first one is $\{\boldsymbol{X}_{ik} : X_{obs} \in (-\infty, median(X_{obs})], k = 1, .., n_i\}$, and the second is $\{\boldsymbol{X}_{ik} : X_{obs} \in (median(X_{obs}), \infty), k = 1, .., n_i\}$. Then, missing values were inserted on $X_{miss}$ as follows: 10% for group one and 30% for the second group.

## 5.1 | Continuous Data

To investigate the type-I error control of the suggested methods, data samples were generated from the model

$$\boldsymbol{X}_{ik} \sim \boldsymbol{F}_i(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_i), \quad i = 1, 2, \quad k = 1, \ldots, n_i,$$

where $\boldsymbol{F}_i(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_i)$ represents a multivariate distribution with expectation vector $\boldsymbol{\mu}_0$ and covariance matrix $\boldsymbol{\Sigma}_i$. Marginal data distributions were generated from different symmetric distributions (normal, double exponential) as well as skewed distributions (lognormal, $\chi^2_{15}$) (similar to Pauly, Brunner, and Konietschke 2015). We used normal copulas to generate the dependency structures of the repeated measurements using the R package **copula** (Yan et al. 2007). For the covariance matrix $\boldsymbol{\Sigma}_i$, we investigated the three following covariance structures

(AR) Setting 1: $\boldsymbol{\Sigma}_i = (\rho^{|l-j|})_{l,j\leq d}, \rho = 0.6$ for $i = 1, 2$,

(CS) Setting 2: $\boldsymbol{\Sigma}_i = \boldsymbol{I}_t$, for $i = 1, 2$,

(TP) Setting 3: $\boldsymbol{\Sigma}_i = (d - |l - j|)_{l,j\leq d}$, for $i = 1, 2$,

representing an autoregressive structure (Setting 1), compound symmetry pattern (Setting 2), and a linear Toeplitz covariance structure (Setting 3). These covariance settings were inspired by the simulations studies in Konietschke et al. (2015), Friedrich, Konietschke, and Pauly (2017), and Umlauft et al. (2019).

**Type-I Error Results**. The type-I error simulation results of the studied procedures for testing the hypotheses of no time effect $H_0^T$, no group × time interaction $H_0^{GT}$, and no group effect $H_0^G$ under the MCAR and MAR frameworks are shown in Tables S.1 and S.12 (MCAR framework) and Tables S.13–S.24 (MAR framework) in the Supporting Information. It can be readily seen that the suggested bootstrap approaches based on $T_W^*, T_A^*$, and $T_M^*$ tend to result in quite accurate type-I error rate control for most hypotheses under symmetric as well as skewed distributions and under MCAR and MAR mechanisms. The type-I error control is surprisingly not affected by less stringent missing mechanisms and the bootstrap tests are robust under fairly large amounts of missing observations. In addition, the data dependency structures do not affect the quality of the approximations. The sample size allocations (balanced vs. unbalanced) slightly impact the type-I error of the tests. On the other hand, the asymptotic ANOVA-type test $T_A$ also shows a quite accurate type-I error control for large sample sizes. However, under small sample sizes, $T_A$ tends to be sensitive to the missing rates. In particular, it exhibits a liberal behavior for larger missing rates. In contrast, the asymptotic Wald test $T_W$ shows an extremely liberal behavior in all considered situations and under all investigated missing mechanisms. A closer look at the type-I error simulation results for all considered settings under the MCAR framework is provided. Compact boxplots factorized in terms of hypotheses of interest or missing rates are given in Figures 1 and 2, respectively. They are based on different sample size settings (4) × different distributions (4) × different missing rates (2) × different covariance structures (3) × different time points (2) × different hypotheses (3). Due to the extreme liberal behavior of the asymptotic Wald test, it has been removed from those plots in order to get a better view of the remaining tests behavior. It can be clearly seen that the asymptotic ANOVA-type test has the less accurate control among all other considered tests. The hypothesis of interest and number of time points obviously impact the quality of the approximations of the asymptotic ANOVA-type test. Furthermore, the type-I error control behavior of the bootstrapped MATS depends on the hypothesis of interest. Consequently, the bootstrapped Wald and ANOVA-type tests are recommended over all considered methods.

In order to cover the *effect of increasing missing rates*, we additionally studied type-I error control for $a = 2$ groups, $d = 4$ time points, $(n_1 = 15, n_2 = 15)$ sample sizes with $r \in \{10\%, 20\%, 30\%, 40\%, 50\%, 60\%\}$ covering missingness in observations ranging from 10% to 60%. Figure 3 and Figures S.1 and S.2 in the Supporting Information summarize type-I error rate control for these settings under symmetric and asymmetric distributions. The results indicate that the asymptotic Wald test $T_W$ tends to be liberal in all considered situations. In particular, it is extremely liberal when testing the hypotheses of no time effect $(H_0^T)$ and no interaction effect $(H_0^{GT})$. In contrast, the asymptotic ANOVA-type test $T_A$ tends to be sensitive to missing rates and hypothesis type. In particular, it exhibits an accurate or liberal behavior for

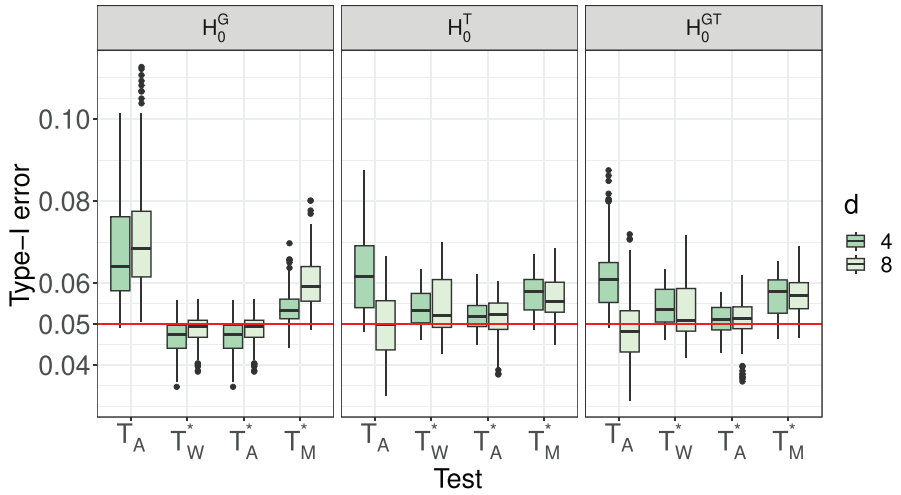



**FIGURE 1** | Type-I error of the asymptotic ANOVA-type test $T_A$ and the bootstrapped tests $T_W^*$, $T_A^*$, and $T_M^*$ based on several hypotheses of interest for varying time points $d \in \{4, 8\}$. Each boxplot summarizes the type-I error results from 96 different simulation scenarios for this hypothesis. For all individual simulations, see the tables in the Supporting Information.

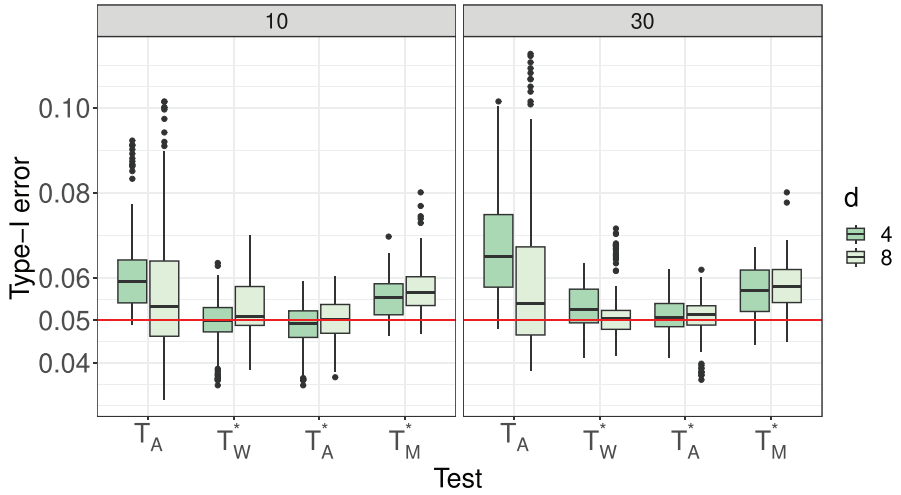



**FIGURE 2** | Type-I error of the asymptotic ANOVA-type test $T_A$ and the bootstrapped tests $T_W^*$, $T_A^*$, and $T_M^*$ based on missing rates $r \in \{10, 30\}$ for varying time points $d \in \{4, 8\}$. Each boxplot summarizes the type-I error results from 144 different simulation scenarios for this missing rate. For all individual simulations, see the tables in the Supporting Information.

small or large missing rates, respectively. Moreover, it shows a quite constant liberal behavior when testing the hypothesis of no group effect ($H_0^G$). Its behavior appears to be independent of the covariance pattern. In contrast, the suggested bootstrap approaches tend to control type-I error rate more accurate over the range of missing rates $r$ for almost all settings.

Further, it was also interesting to discover the type-I error rate control of the tests under *similar attributes to the data sets of the fluvoxamine and skin disorder clinical trials*. The data sets reflect large sample sizes and a small or moderate amount of missing values from either a one-sample repeated measurements design or a two-way layout design. Simulation results for the type-I error rate of the studied procedures for $(n = 299, d = 3)$ and $(n_1 = 88, n_2 = 84, d = 3)$ sample sizes are presented in Table S.25 and Table S.26 in the Supporting Information, respectively. It can be seen that most tests are robust under both settings and control type-I error rate accurately. Only the Wald-type tests exhibit a liberal behavior for some of the lognormal settings.

In addition, we investigated the *impact of increasing the number of groups* on type-I error rate control. Starting with two groups, we systematically expanded the number of groups up to 12. Our observations were generated from a lognormal distribution, with sample sizes set in an unbalanced manner as follows: $n = (n_1, \ldots, n_{12}) = (10, 10, 15, 20, 35, 25, 15, 30, 20, 35, 20, 15)$. We considered two time points ($d \in \{4, 8\}$), an autoregressive covariance structure (Setting 1), chosen for its similarity to other structures, and a missing rate of 30%. This simulation setup was inspired by a previous study on repeated measures designs with a potentially large number of groups by Sattler (2021). Figure 5 and Figure S.3 in the Supporting Information, corresponding to hypotheses $H_0^G$ and $H_0^{GT}$, respectively, illustrate that the asymptotic Wald test $T_W$ consistently demonstrates a liberal behavior across all

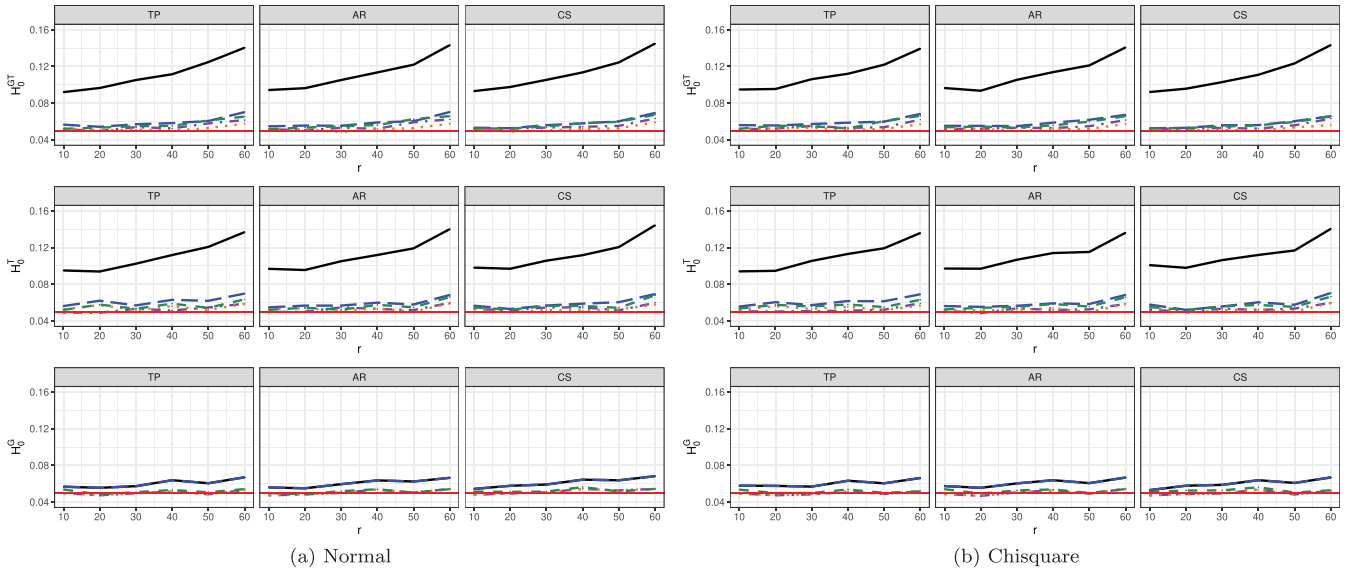



**FIGURE 3** | Type-I error simulation results ($\alpha = 0.05$) of the tests $T_W$ (– – – – –), $T_A$ (——), $T_W^*$ (–·–·–), $T_A^*$ (···), and $T_M^*$ (– – –) under different covariance structures with sample sizes $(n_1, n_2) = (15, 15)$ and $d = 4$ for varying percentages of MCAR data $r \in \{10\%, 20\%, 30\%, 40\%, 50\%, 60\%\}$ with observations generated from (a) a normal and (b) a $\chi^2_{15}$ distribution, respectively.

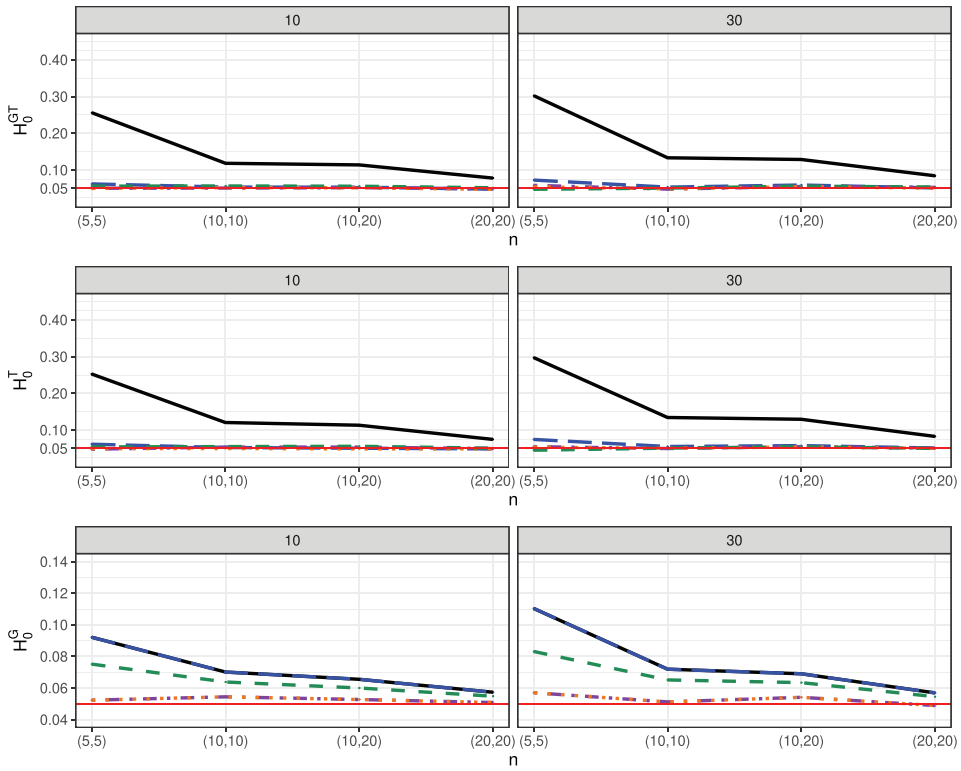



**FIGURE 4** | Type-I error simulation results ($\alpha = 0.05$) of the tests $T_W$ (– – – – –), $T_A$ (——), $T_W^*$ (–·–·–), $T_A^*$ (···), and $T_M^*$ (– – –) for ordinal data under MCAR framework with sample sizes $(n_1, n_2) = \{(5, 5), (10, 10), (10, 20), (20, 20)\}$, and $d = 4$.

scenarios. However, the asymptotic ANOVA-type test $T_A$ shows sensitivity to the number of time points $d$, displaying more accurate behavior for a small number of time points ($d = 4$). In contrast, the suggested bootstrap ANOVA-type test $T_A^*$ tends to provide more accurate control of the type-I error rate across almost all settings and among all considered tests. Moreover, it is noteworthy that the simulation results reveal a consistent pattern: When considering scenarios under the hypothesis $H_0^G$, the behavior of the WTS $T_W$ and its respective wild bootstrap version $T_W^*$ deteriorates with an increasing number of groups, while the opposite is observed for the two ANOVA-type approaches.

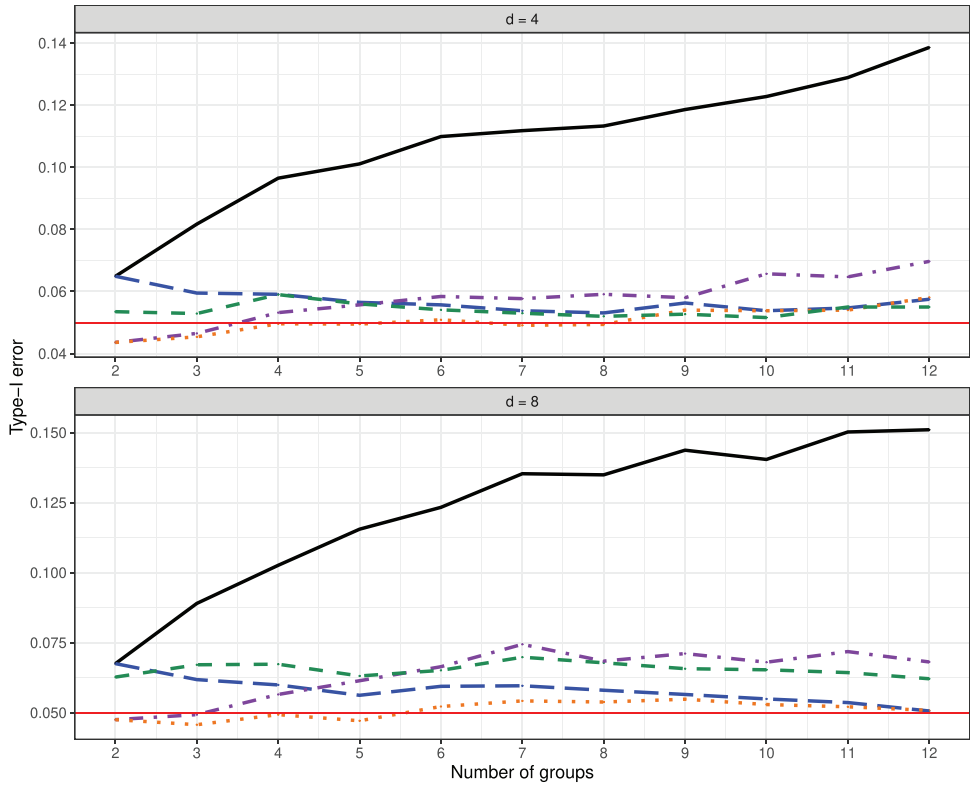



**FIGURE 5** | Type-I error simulation results ($\alpha = 0.05$) of the tests $T_W$ (– – – – –), $T_A$ (— —), $T_W^*$ (– · – ·), $T_A^*$ (· · · ·), and $T_M^*$ (– – – –) for increasing number of groups under the MCAR framework with sample sizes $n = (n_1, \cdots, n_{12}) = (10, 10, 15, 20, 35, 25, 15, 30, 20, 35, 20, 15)$, $d \in \{4, 8\}$ with observations generated from a lognormal distribution and under the hypothesis $H_0^G$.

## 5.2 | Ordinal Data

In order to address all the goals outlined above, we simulated ordinal data. The observations were simulated similar to Brunner and Langer (2000) as follows:

$$\boldsymbol{X}_{ijk} = int\left(4\frac{cZ_{ik} + Y_{ijk}}{c+1}\right) + 1, \quad i = 1, \ldots, a, \quad k = 1, \ldots, n_i, \quad j = 1, \ldots, d,$$

where $Z_{ik}$ and $Y_{ijk}$ are independently uniformly distributed in the interval [0,1], $c > 0$ is a constant, and $int(x)$ indicates the integer part of $x$. The elements of $\boldsymbol{X}_{ijk}$ take values between 1 and 4. The correlation between $\boldsymbol{X}_{ij\}k}$ and $\boldsymbol{X}_{ijk}$ is determined by the choice of the constant $c$. In our simulation study, we considered $c = 1$ assuring a compound symmetric covariance structure. We considered $a = 2$ groups and $d \in \{4, 8\}$ dimensions, and the same sample sizes as in the continuous data settings.

The type-I error results of the considered methods under MCAR and MAR frameworks are summarized in Figure 4 and Figures S.4–S.6 in the Supporting Information. It can be seen that the asymptotic Wald test is too liberal in most situations and under all considered hypotheses. The ANOVA-type test of Brunner, Munzel, and Puri (1999) shows much better behavior. In contrast, the type-I error control for the bootstrap-based procedures is the best, particularly for the bootstrapped Wald and ANOVA-type tests, which are also less affected by an increased missing rate or strict MAR assumptions.

## 5.3 | Power

In order to assess the empirical power of all studied methods, we considered a one-sample repeated measures design with $d = 4$ repeated measures, sample size $n = 15$, and covariance structures as given in Settings $1 - 3$ for various distributions. Data were generated by

$$[\boldsymbol{X}_{1k} \sim \boldsymbol{F}_1(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_1)] + \boldsymbol{\mu}_1, \quad k = 1, \ldots, 15,$$

where, we were interested in detecting two specific alternatives

- Alternative 1: $\boldsymbol{\mu}_1 = (0, 0, \zeta, \zeta)$,
- Alternative 2: $\boldsymbol{\mu}_1 = (0, 0, 0, \zeta)$,

for varying shift parameter $\zeta \in \{0, 0.5, 1, 1.5, 2, 2.5, 3\}$.

The power analysis results of the considered methods under the MCAR framework for several distributions, involving various covariance settings for detecting Alternative 1 are summarized in Figures 6 and 7 (missing rate $r = 30\%$) and Figures S.7 and S.8 in the Supporting Information ($r = 10\%$). The simulation results for investigating Alternative 2 are displayed in Figures S.9 and S.12 in the Supporting Information. The power analysis results of the considered methods under the respective MAR framework are summarized in Figures S.13–S.16 (MAR1 scenario) and Figures S.17–S.20 (MAR2 scenario).

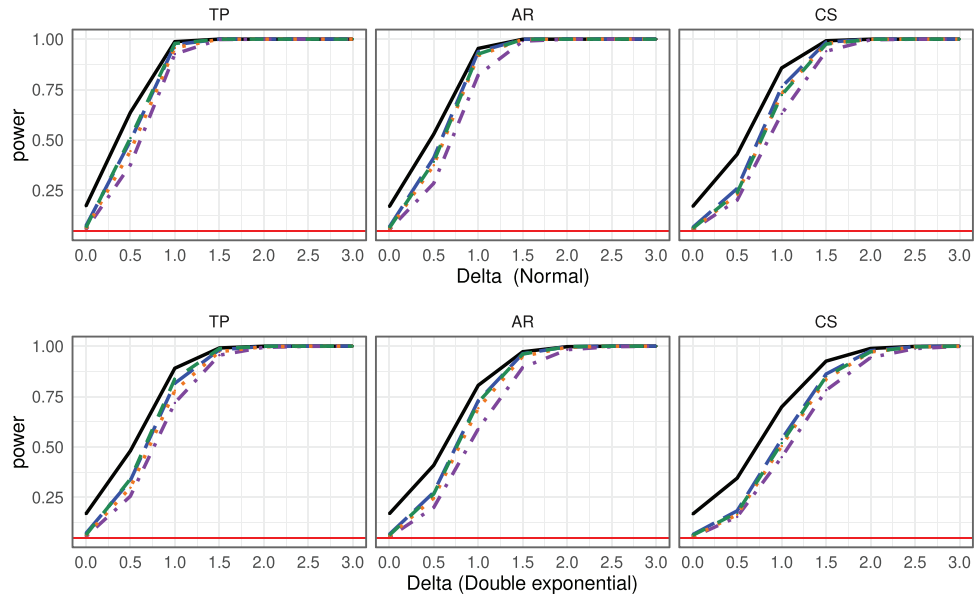



**FIGURE 6** | Power simulation results of the tests $T_W$ (– – – – –), $T_A$ (——), $T_W^*$ (– · –), $T_A^*$ (· · ·), and $T_M^*$ (– – –) under different covariance structures with sample size $n = 15$ and $d = 4$ under alternative 1, under the MCAR framework with missing rate $r = 30\%$ with observations generated from a normal (upper row) and a double exponential (bottom row) distribution, respectively.

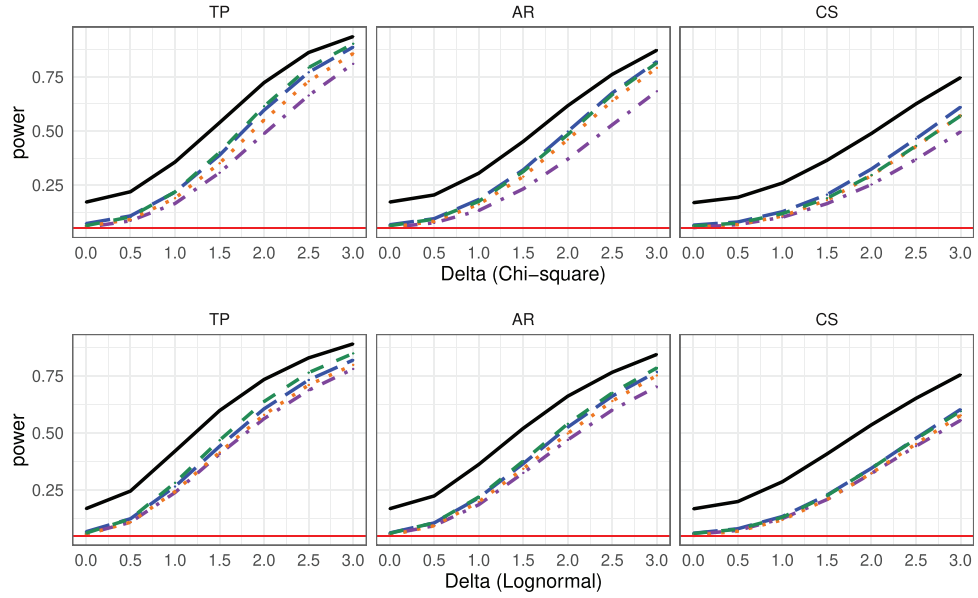



**FIGURE 7** | Power simulation results of the tests $T_W$ (– – – – –), $T_A$ (——), $T_W^*$ (– · –), $T_A^*$ (· · ·), and $T_M^*$ (– – –) under different covariance structures with sample size $n = 15$ and $d = 4$ under alternative 1, under the MCAR framework with missing rate $r = 30\%$ with observations generated from a $\chi^2_{15}$ (upper row) and a lognormal (bottom row) distribution, respectively.

As the asymptotic Wald test is too liberal compared to the other studied methods, its power function is larger. Moreover, the bootstrapped Wald test exhibits the lowest power behavior, while the bootstrapped MATS and ANOVA-type have a quite similar power behavior as the Brunner, Munzel, and Puri (1999) ANOVA-type test. The differences between the procedures is less pronounced under the MAR framework.

To sum up, we recommend the bootstrap ATS. It exhibits the overall best type-I error control combined with a good power behavior and needs the less stringent assumptions for application.

## 5.4 | Additional Comparative Simulation Results

In this section, we present additional comparative simulation results for other wild bootstrap weights (5.4.1) and a comparison with recent procedures that infer hypotheses in terms of nonparametric effects (5.4.2).

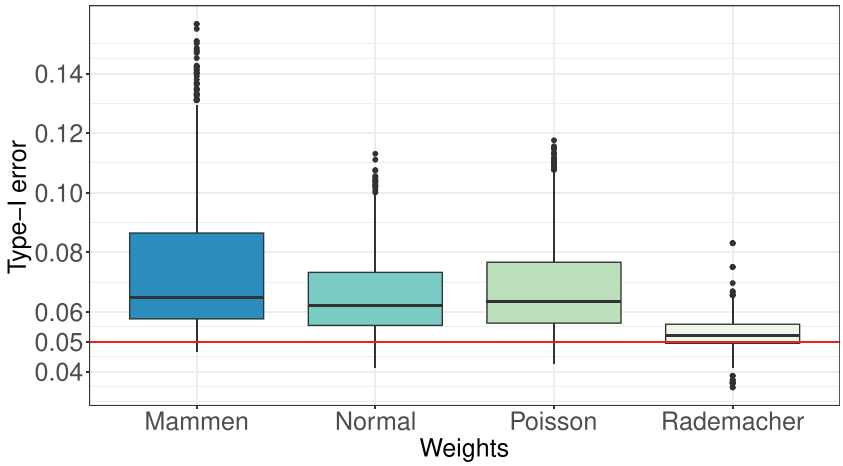



**FIGURE 8** | Type-I error for three different wild bootstrap versions of our bootstrapped tests $T_W^*$, $T_A^*$, and $T_M^*$. Each boxplot summarizes the type-I error results from 288 different simulation scenarios.

#### 5.4.1 | Considering Different Weights in the Wild Bootstrap Method

Here, we investigate the impact of different wild bootstrap weights $W_{ik}$ on the performance of our methods, aiming to identify any potentially "optimal" weights that may enhance our tests' effectiveness. We examine the following weight distributions as previously proposed in the literature in other contexts:

- Rademacher weights as above: $P(W_{ik} = -1) = P(W_{ik} = 1) = 1/2$, see also Liu (1988) and Friedrich, Konietschke, and Pauly (2017).
- Mammen weights: $P(W_{ik} = -(\sqrt{5}-1)/2) = (\sqrt{5}+1)/(2\sqrt{5}), P(W_{ik} = (\sqrt{5}+1)/2) = 1-(\sqrt{5}+1)/(2\sqrt{5})$, see Mammen (1993).
- Normal weights: $W_{ik} \sim N(0,1)$, see Lin (1997).
- Centered Poisson weights: $W_{ik} \sim Po(1) - 1$, see Beyersmann, Termini, and Pauly (2013).

Normal and centered Poisson weights have been recommended in the survival setting (the above references) while Mammen weights were proposed in the context of high-dimensional linear models (Mammen 1993). We compare the performance of these weight distributions with our suggested wild bootstrap tests and summarize the results in Figure 8. Notably, we find that Rademacher weights exhibit the best type-I error control among the four different approaches. Consequently, we recommend the adoption of Rademacher weights due to their superior performance compared to the other three weight distributions examined.

#### 5.4.2 | Comparison With Further Alternative Approaches

This section provides a comparative analysis between our proposed methods and recent approaches based on asymptotics rather than resampling techniques (Rubarth, Pauly, and Konietschke 2022; Rubarth et al. 2022). These alternative approaches target specific scenarios: The former paper addresses repeated measures designs, which can be regarded as a special case of our design, while the latter deals with clustered data. Both our methods and the methods proposed by Rubarth, Pauly, and Konietschke (2022) and Rubarth et al. (2022) use all-available data and are valid under the MCAR mechanism. However, they differ in their scope: While Rubarth, Pauly, and Konietschke (2022) and Rubarth et al. (2022) concentrate on hypotheses formulated in terms of relative marginal effects, our proposed procedures are constrained to testing null hypotheses formulated in terms of the equality of the distribution functions. Consequently, the two approaches are not directly comparable. However, one can show that our null hypothesis implies the null hypothesis formulated in terms of relative marginal effects (Brunner, Bathke, and Konietschke 2018). This enables a rough comparison between the behaviors of these approaches. Given the distinct setups studied in the aforementioned papers, we conducted two separate simulation studies aligning with the statistical model designs outlined in each paper.

(1) **Alternative methods developed for repeated measures designs:**
We investigate the approaches proposed by Rubarth, Pauly, and Konietschke (2022), which introduce nonparametric methods for analyzing repeated measures designs with missing data. They developed various test procedures, including global and multiple contrast tests, to test the null hypothesis $H_0^{p^*} : \boldsymbol{C}\boldsymbol{p}^* = \boldsymbol{0}$, where the unweighted relative marginal effect are collected in the vector $\boldsymbol{p}^* = (p_1^*, \ldots, p_d^*)^T$. Here, $p_i^* = \int G(x)dF_i(x)$ denotes the unweighted relative effects, while $G(x) = d^{-1}\sum_{i=1}^{d} F_i(x)$ is the unweighted average of all distribution functions in the experiment. They developed generalized versions of the quadratic form tests proposed by Domhof, Brunner, and Osgood (2002) to test the less stringent hypothesis $H_0^{p^*} : \{\boldsymbol{C}\boldsymbol{p}^* = \boldsymbol{0}\}$. We note that $H_0^{p^*}$ implies our null hypothesis $H_0 : \boldsymbol{C}\boldsymbol{F} = \boldsymbol{0}$.
Based on their simulation study and recommendations, we compare our approaches with their generalized Domhof, Brunner, and Osgood (2002) statistic (A1), their newly introduced ATS (A2) with its distribution approximated via the Greenhouse–Gaisser (Box 1954) method, as well

**TABLE 1** | Simulation results for type-I error level ($\alpha = 0.05$) for normal distribution and different percentages $r$ in case of the MCAR mechanism.

| $d$ | Cov | $n$ | $r = 10$ | | | | | | $r = 30$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | WBtstrp | | | Alternatives | | | WBtstrp | | | Alternatives | | |
| | | | $T^*_W$ | $T^*_A$ | $T^*_M$ | $A1$ | $A2$ | $M$ | $T^*_W$ | $T^*_A$ | $T^*_M$ | $A1$ | $A2$ | $M$ |
| 4 | AR | 10 | 5 | 5.2 | 5.7 | 7.5 | 5 | 10.3 | 6.2 | 5.9 | 7.4 | 9.6 | 7.1 | 15 |
| | | 15 | 5.4 | 5.3 | 5.8 | 6.8 | 5.3 | 8.7 | 5.8 | 5.7 | 6.3 | 8.1 | 6.4 | 11.1 |
| | | 20 | 4.9 | 4.9 | 5.2 | 5.9 | 4.9 | 7.1 | 4.6 | 5 | 5.3 | 6.7 | 5.5 | 8.9 |
| | | 30 | 5.6 | 5.6 | 5.7 | 6.2 | 5.6 | 6.7 | 5.5 | 5.6 | 5.8 | 6.7 | 6 | 7.8 |
| | CS | 10 | 5 | 5 | 5.9 | 7.3 | 4.7 | 10.6 | 6.3 | 5.8 | 6.9 | 9.3 | 6.3 | 14.7 |
| | | 15 | 5.1 | 5 | 5.3 | 6 | 4.5 | 8.1 | 5.3 | 5 | 5.7 | 7.2 | 5.5 | 10.4 |
| | | 20 | 5.4 | 5.5 | 5.6 | 6.3 | 5.2 | 7.2 | 5.4 | 5.8 | 6 | 7.3 | 6 | 9.5 |
| | | 30 | 5.4 | 5.3 | 5.6 | 5.7 | 5.1 | 6.7 | 5.4 | 5.3 | 5.8 | 6.3 | 5.5 | 8 |
| | TP | 10 | 5.1 | 6.1 | 6.7 | 8.9 | 6.5 | 11 | 5.9 | 6 | 7.2 | 9.8 | 7.2 | 14.3 |
| | | 15 | 5.3 | 5.6 | 5.8 | 7.5 | 6.2 | 8.5 | 5.3 | 5.4 | 5.8 | 8.1 | 6.6 | 10.6 |
| | | 20 | 4.9 | 5.8 | 5.8 | 7.1 | 6.2 | 7.5 | 5.8 | 6 | 6.4 | 7.9 | 6.9 | 9.4 |
| | | 30 | 4.8 | 5.2 | 5.4 | 6.3 | 5.5 | 6.3 | 5.4 | 5.4 | 5.6 | 6.4 | 5.8 | 7.6 |

as their maximum-type statistic (M). Since their paper focuses on testing the hypotheses of no time effect, we solely investigated the hypothesis $H_0^T$. Motivated from above, we considered a one-sample repeated measures design with two different time points $d \in \{4, 8\}$, sample sizes $n \in \{10, 15, 20, 30\}$, and covariance structures as specified in Settings 1–3 for various distributions. The simulation results are provided in Tables S.27–S.30 in the Supporting Information. A glimpse of the analysis results for the specific choice of normally distributed data is presented in Table 1. The results reveal that our resampling methods are more favorable with respect to type-I error control. In fact, the maximum-type testing procedure (M) tends to be liberal across nearly all considered scenarios, while in most scenarios with smaller sample sizes, the other two approaches (A1 and A2) also tend to be liberal. In contrast, our newly proposed methods perform well. A likely reason is that their methods were developed for testing the more complex null hypothesis formulated in terms of effect sizes. They need to estimate way more parameters and therefore they are more liberal in small samples.

(2) **Alternative methods developed for factorial clustered data designs:**
Here, we investigate the methods proposed by Rubarth et al. (2022), which addresses factorial repeated measures designs with clustered data, extending aforementioned research by Rubarth, Pauly, and Konietschke (2022). Similar to Rubarth, Pauly, and Konietschke (2022), they propose quadratic- and maximum-type testing procedures for testing the null hypothesis $H_0^P$. Rubarth et al. (2022) introduce a WTS, an ATS, and a maximum-type statistic. Given the liberal behavior of the WTS in small or moderate sample size scenarios, they recommend the other two, denoted as $A_{cl}$ and $M_{cl}$, respectively, in our simulation results. As their model can handle clustered data structure, our model is contained within theirs as a special case. To allow comparison between their methods and ours, we set the number of possibly dependent replicates of any subject in any group at any time to be $m_{ijk} = 1$. We examined similar simulation scenarios to those in our study. Specifically, we adopted a two-way layout design with $a = 2$ independent groups and $d = 3$. We explored both balanced and unbalanced situations with sample size vectors $(n_1, n_2) \in \{(5, 5), (10, 10), (20, 20), (10, 20), (15, 15), (30, 30)\}$, alongside the three hypotheses: $H_0^G$, $H_0^T$, and $H_0^{GT}$ under the MCAR framework. Our findings, presented in Figure 9 and Tables S.31–S.38 in the Supporting Information, indicate that the methods $A_{cl}$ and $M_{cl}$, developed by Rubarth et al. (2022), have problems in keeping the type-I error rate in our situation. In particular, they exhibit a liberal behavior across various scenarios. In contrast, our proposed methods, especially the bootstrapped version of the ANOVA type-test $T^*_A$, demonstrate much more favorable behavior. Again, a likely reason is that their methods have been developed for a more complex setting.

## 6 | Application to Empirical Data

### 6.1 | The Fluvoxamine Trial

In this section, we re-examine a clinical fluvoxamine study. It has already been analyzed by Molenberghs and Lesaffre (1994), Molenberghs, Kenward, and Lesaffre (1997), Van Steen et al. (2001), Jansen et al. (2003), Molenberghs and Verbeke (2004), and Molenberghs and Kenward (2007). The study has been performed to establish the profile of fluvoxamine in ambulatory clinical psychiatric operations. Hereby, 315 patients who suffer from depression, panic disorder, and/or obsessive-compulsive disorder were scored every 2 weeks over 6 weeks of treatment ($d = 3$). At each clinical visit, scores for both *side effect* and *therapeutic effect* scales, which are based on about 20 psychiatric symptoms were recorded. The side effect scale ranges from 1 to 4. The lower the score, the better the clinical record. For

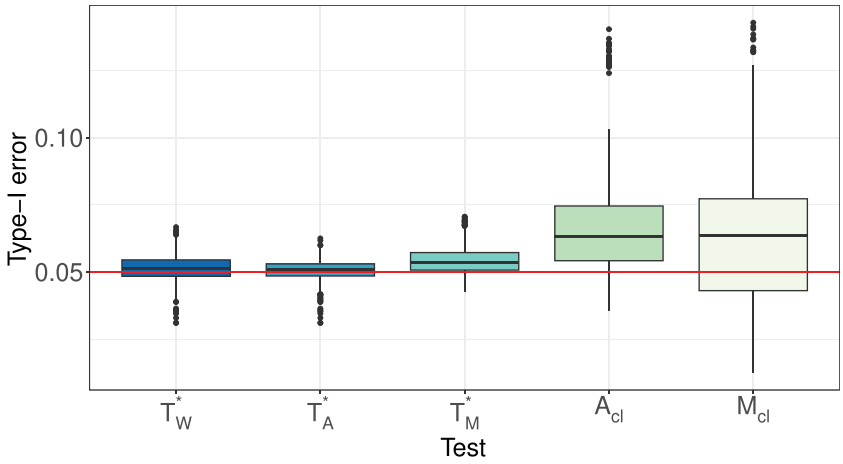
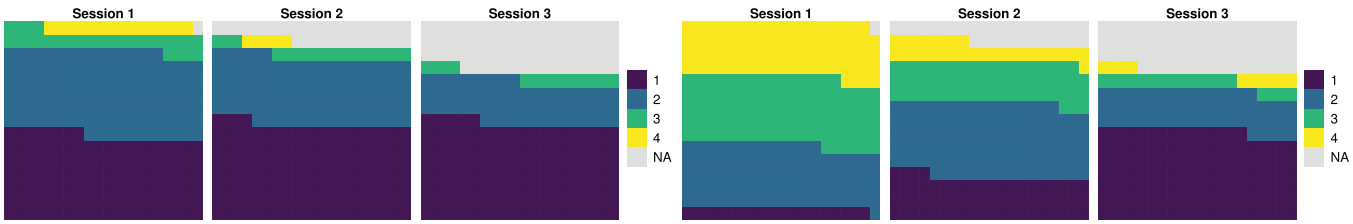



**FIGURE 9** | Type-I error for three different wild bootstrap versions of our bootstrapped tests $T_W^*, T_A^*$, and $T_M^*$ and the two asymptotic tests of Rubarth et al. (2022) $A_{cl}$ and $M_{cl}$. Each boxplot summarizes the type-I error results from 432 different simulation scenarios.




(a) **Side effect** (b) **Therapeutic effect**

**FIGURE 10** | Frequencies of the side effect and therapeutic effect scores observed in the fluvoxamine trial.

example, score 1 stands for "no side effect" while score 4 indicates that "the side effect surpasses the therapeutic effect." Similarly, the therapy effect is a four-category ordinal scale: (1) "no improvement or worsening"; (2) "minimal improvement, not changing functionality"; (3) "moderate improvement, partial disappearance of symptoms"; and (4) "important improvement, almost disappearance of symptoms." The higher the score, the better the patients health.

Several patients missed the recording of their measurements in some sessions, which leaded to a large amount of missing values. A closer look to our data shows that, from the total of 315 initially recruited patients, 14 patients dropped off, and two were excluded from the analysis due to their nonmonotone missing pattern. This leaves us with a total of 299 patients who have at least one measurement. Among them, 242 patients have complete observations, 31 were scored only on the first session, and 44 were scored on both Session 1 and Session 3. Waffle plots representing the distributions of the side effect and therapeutic effect among the three sessions are shown in Figure 10. We aim to test the hypotheses whether side effect or therapeutic effect scores are significantly different between the three sessions for patients with psychiatric disorder. To this end, we applied all considered testing methods; asymptotic Wald and ANOVA-type tests ($T_W, T_A$), and the bootstrap procedures ($T_W^*$, $T_A^*$, $T_M^*$) to detect the null hypothesis $H_0^T : \{\boldsymbol{CF} = \boldsymbol{0}\}$. The results are summarized in Table 2. It can be seen that all tests indicate a significant difference between the therapeutic effect scores as well as the side effect scores of the three sessions (two-sided

**TABLE 2** | $p$-Values of the fluvoxamine study.

| Effect | $T_W$ | $T_A$ | $T_W^*$ | $T_A^*$ | $T_M^*$ |
|---|---|---|---|---|---|
| Therapeutic effect | $2.3e^{-79}$ | $2.2e^{-86}$ | 0 | 0 | 0 |
| Side effect | $2.3e^{-10}$ | $4.1e^{-12}$ | 0 | 0 | 0 |

$p$-value $< 0.00001$). Moreover, we conclude that the clinical outcome of the patients significantly improves after three sessions. These findings coincide with that in Molenberghs and Kenward (2007).

## 6.2 | The Skin Disorder Trial

Here, we study data from a randomized, multicenter, parallel group study for treating a skin condition. Treating skin conditions can be tough, thus the goal of the study was to assess the severe rate of the skin condition over time and to compare the efficiency and safety of two continuous therapy treatments drug and placebo. Patients were randomly assigned to drug or placebo therapy treatment. Prior to treatment, patients were assessed to determine the initial severity of the skin condition (moderate or severe). At three follow-up visits, the treatment outcome was measured according to a five-point ordinal response scale that assess the extent of improvement (1 = rapidly improving, 2 = slowly improving, 3 = stable, 4 = slowly worsening, 5 = rapidly worsening). The study consists of 88 and 84 subjects allocated to

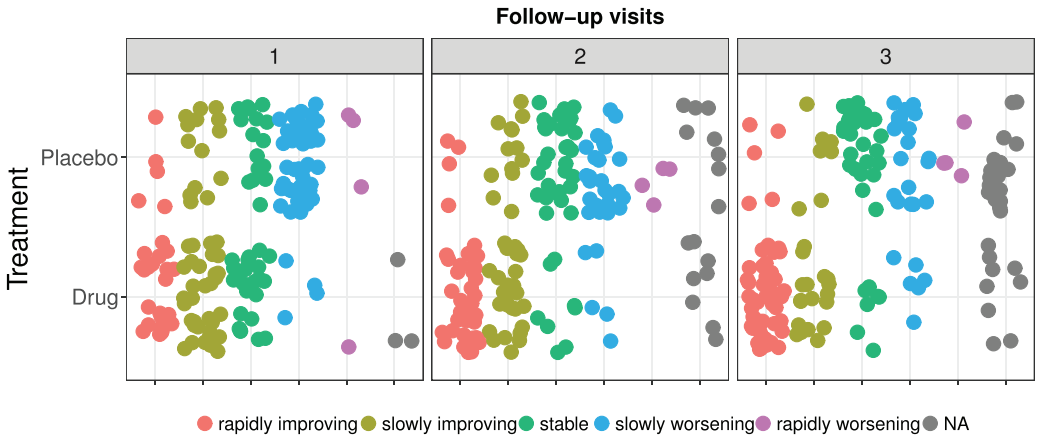



**FIGURE 11** | Frequencies of observed patients treatment outcome in the skin disorder trial.

**TABLE 3** | *p*-Values of the skin disorder trial.

| Hypothesis | $T_W$ | $T_A$ | $T_W^*$ | $T_A^*$ | $T_M^*$ |
|---|---|---|---|---|---|
| Group | 0 | 0 | 0 | 0 | 0 |
| Time | 0 | 0 | 0 | 0 | 0 |
| Group × Time | 0.032 | 0.016 | 0.026 | 0.01 | 0.009 |

the active treatment group and the placebo group, respectively. And, the proportion of missing observations is around 30%. The distribution of patients improvement across the treatment groups and the follow-up visits is displayed in Figure 11. The study is described in full detail in Landis et al. (1988) and is published in Davis (2002). There is not enough information available from the data source regarding the missingness mechanism. It is likely that the missingness is not at random, considering the nature of the study. Nonetheless, for the sake of illustration, we continue with our analysis.

Similar to the fluvoxamine study above, we applied all asymptotic and bootstrap procedures to infer the following null hypotheses: "no group effect," "no time effect," and "no group × time effect." The results are summarized in Table 3: All approaches reject the null hypothesis of no group effect, the null hypothesis of no time effect, and the null hypothesis of no treatment group × time interaction. This implies that the clinical outcome of the patients significantly improves with time and this progression is significantly different between the two treatment groups, drug and placebo. Therefore, the data are further analyzed and split by the factor initial severity and the analysis is replicated separately for each baseline severity level (moderate or severe). The results are provided in Table 4.

It can be seen from Table 4 that all approaches detect a significant group effect as well as a significant time effect in both moderate and severe groups. In contrast, a significant group × time interaction effect arises only in the moderate severity group. This indicates that the change in clinical outcomes of patients of moderate severity varies over time depending on treatment group membership. All five approaches share the previous findings.

## 7 | Summary

Multigroup repeated measures design with ordinal or skewed observations are quite common. If the observations are additionally subject to missing values existing methods for testing null hypotheses in terms of distribution function may be either liberal (WTSs) or run (asymptotically) on a wrong type-I error level (ATS; Domhof, Brunner, and Osgood 2002). To this end, we investigated three alternatives based on resampling. We proved their asymptotic validity and analyzed their small sample behavior regarding type-I error control and power in extensive simulations. Under all of the five considered methods, an ATS with critical values calculated by means of a Wild bootstrap approach exhibits the best behavior and is recommended.

In the future, we will include the present methodology into the R package nparLD (Noguchi et al. 2012). Moreover, we plan to extend our investigations to general MANOVA settings, for example, extending the results of Dobler, Friedrich, and Pauly (2020) to the situation with missing values.

**TABLE 4** | *p*-Values of the split skin disorder trial based on initial severity.

| | Initial severity = Moderate | | | | | Initial severity = Severe | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Hypothesis** | $T_W$ | $T_A$ | $T_W^*$ | $T_A^*$ | $T_M^*$ | $T_W$ | $T_A$ | $T_W^*$ | $T_A^*$ | $T_M^*$ |
| Group | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Time | 0.003 | 0.004 | 0.003 | 0.002 | 0.002 | 0 | 0 | 0.001 | 0 | 0 |
| Group × Time | 0.041 | 0.032 | 0.043 | 0.027 | 0.029 | 0.423 | 0.318 | 0.447 | 0.338 | 0.335 |

## Acknowledgments

Lubna Amro and Markus Pauly acknowledge the support of the German Research Foundation (DFG) (No. PA 2409/3-2). Frank Konietschke's work was also supported by the German Research Foundation (DFG) (No. KO 4680/3-2). Lubna Amro also acknowledges the support of the project "From Prediction to Agile Interventions in the Social Sciences (FAIR)," which receives funding from the program "Profilbildung 2020," an initiative of the Ministry of Culture and Science of the State of Northrhine Westphalia. The sole responsibility for the content of this publication lies with the authors.

Open access funding enabled and organized by Projekt DEAL.


## Conflicts of Interest

The authors have declared no conflicts of interest.

## Data Availability Statement

The data that support the findings of this study are available in the Supporting Information of this article.

## Open Research Badges

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to computational complexity.

## References


Akritas, M. G. 2011. "Nonparametric Models for ANOVA and ANCOVA Designs." In, edited by M. Lovric, 964–968. *International Encyclopedia of Statistical Science*. Berlin, Heidelberg: Springer.

Akritas, M. G., and S. F. Arnold. 1994. "Fully Nonparametric Hypotheses for Factorial Designs I: Multivariate Repeated Measures Designs." *Journal of the American Statistical Association* 89, no. 425: 336–343.

Akritas, M. G., and E. Brunner. 1997. "A Unified Approach to Rank Tests for Mixed Models." *Journal of Statistical Planning and Inference* 61, no. 2: 249–277.

Akritas, M. G., J. Kuha, and D. W. Osgood. 2002. "A Nonparametric Approach to Matched Pairs With Missing Data." *Sociological Methods & Research* 30, no. 3: 425–454.

Amro, L., M. Pauly, and B. Ramosaj. 2019. "Asymptotic Based Bootstrap Approach for Matched Pairs With Missingness in a Single-arm." *Biometrical Journal* 63, no. 7: 1389–1405.

Arnau, J., R. Bono, M. J. Blanca, and R. Bendayan. 2012. "Using the Linear Mixed Model to Analyze Nonnormal Data Distributions in Longitudinal Designs." *Behavior research methods* 44, no. 4: 1224–1238.

Bartlett, M. S. 1939. "A Note on Tests of Significance in Multivariate Analysis." *Mathematical Proceedings of the Cambridge Philosophical Society* 35, no. 2: 180–185.

Bathke, A. C., S. Friedrich, M. Pauly, et al. 2018. "Testing Mean Differences Among Groups: Multivariate and Repeated Measures Analysis With Minimal Assumptions." *Multivariate Behavioral Research* 53, no. 3: 348–359.

Beyersmann, J., S. D. Termini, and M. Pauly. 2013. "Weak Convergence of the Wild Bootstrap for the Aalen–Johansen Estimator of the Cumulative Incidence Function of a Competing Risk." *Scandinavian Journal of Statistics* 40, no. 3: 387–402.

Biederman, L. A., T. W. Boutton, and S. G. Whisenant. 2008. "Nematode Community Development Early in Ecological Restoration: The Role of Organic Amendments." *Soil Biology and Biochemistry* 40, no. 9: 2366–2374.

Bourlier, V., A. Zakaroff-Girard, A. Miranville, et al. 2008. "Remodeling Phenotype of Human Subcutaneous Adipose Tissue Macrophages." *Circulation* 117, no. 6: 806.

Box, G. E. 1954. "Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems, I. Effect of Inequality of Variance in the One-Way Classification." *Annals of Mathematical Statistics* 25, no. 2: 290–302.

Brombin, C., E. Midena, and L. Salmaso. 2013. "Robust Non-Parametric Tests for Complex-Repeated Measures Problems in Ophthalmology." *Statistical Methods in Medical Research* 22, no. 6: 643–660.

Brunner, E. 2001. "Asymptotic and Approximate Analysis of Repeated Measures Designs Under Heteroscedasticity." In *Mathematical Statistics With Applications in Biometry*. eds. J. Kunert and G. Trenkler (Koeln: Josef Eul Verlag, Lohmar), 313–325.

Brunner, E., A. C. Bathke, and F. Konietschke. 2018. *Rank and Pseudo-Rank Procedures for Independent Observations in Factorial Designs*. Springer Nature Switzerland AG, Switzerland: Springer Series in Statistics.

Brunner, E., F. Konietschke, M. Pauly, and M. L. 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, no. 5: 1463–1485.

Brunner, E., and F. Langer. 2000. "Nonparametric Analysis of Ordered Categorical Data in Designs With Longitudinal Observations and Small Sample Sizes." *Biometrical Journal* 42, no. 6: 663–675.

Brunner, E., U. Munzel, and M. L. Puri. 1999. "Rank-Score Tests in Factorial Designs With Repeated Measures." *Journal of Multivariate Analysis* 70, no. 2: 286–317.

Brunner, E., and M. L. Puri. 2001. "Nonparametric Methods in Factorial Designs." *Statistical Papers* 42, no. 1: 1–52.

Davidson, R., and E. Flachaire. 2008. "The Wild Bootstrap, Tamed at Last." *Journal of Econometrics* 146, no. 1: 162–169.

Davis, C. S. 2002. *Statistical Methods for the Analysis of Repeated Measurements*. Springer New York, NY: Springer Science & Business Media.

Dobler, D., S. Friedrich, and M. Pauly. 2020. "Nonparametric MANOVA in Meaningful Effects." *Annals of the Institute of Statistical Mathematics* 72: 997–1022.

Dobler, D., M. Pauly, and T. Scheike. 2019. "Confidence Bands for Multiplicative Hazards Models: Flexible Resampling Approaches." *Biometrics* 75, no. 3: 906–916.

Domhof, S., E. Brunner, and D. W. Osgood. 2002. "Rank Procedures for Repeated Measures With Missing Values." *Sociological Methods & Research* 30, no. 3: 367–393.

Erceg-Hurn, D. M., and V. M. Mirosevich. 2008. "Modern Robust Statistical Methods: An Easy Way to Maximize the Accuracy and Power of Your Research." *American Psychologist* 63, no. 7: 591.

Fitzmaurice, G. M., N. M. Laird, and J. H. Ware. 2012. *Applied Longitudinal Analysis*. Vol 998. Hobokon, New Jersey: John Wiley & Sons.

Fong, Y., Y. Huang, M. P. Lemos, and M. J. Mcelrath. 2018. "Rank-Based Two-Sample Tests for Paired Data With Missing Values." *Biostatistics* 19, no. 3: 281–294.

Friedrich, S., E. Brunner, and M. Pauly. 2017. "Permuting Longitudinal Data in Spite of the Dependencies." *Journal of Multivariate Analysis* 153: 255–265.

Friedrich, S., F. Konietschke, and M. Pauly. 2017. "A Wild Bootstrap Approach for Nonparametric Repeated Measurements." *Computational Statistics & Data Analysis* 113: 38–52.

Friedrich, S., and M. Pauly. 2018. "MATS: Inference for Potentially Singular and Heteroscedastic MANOVA." *Journal of Multivariate Analysis* 165: 166–179.

Gao, X. 2007. "A Nonparametric Procedure for the Two-Factor Mixed Model With Missing Data." *Biometrical journal* 49, no. 5: 774–788.

Jansen, I., G. Molenberghs, M. Aerts, H. Thijs, and K. Van Steen. 2003. "A Local Influence Approach Applied to Binary Data From a Psychiatric Study." *Biometrics* 59, no. 2: 410–419.

Johnson, R. A., and D. W. Wichern. 2007. *Applied Multivariate Statistical Analysis*. Vol 6. Upper Saddle River, NJ: Prentice Hall.

Konietschke, F., A. C. Bathke, S. W. Harrar, and M. Pauly. 2015. "Parametric and Nonparametric Bootstrap Methods for General MANOVA." *Journal of Multivariate Analysis* 140: 291–301.

Konietschke, F., S. W. Harrar, K. Lange, and E. Brunner. 2012. "Ranking Procedures for Matched Pairs With Missing Data–Asymptotic Theory and a Small Sample Approximation." *Computational Statistics & Data Analysis* 56, no. 5: 1090–1102.

Krishnamoorthy, K., and F. Lu. 2010. "A Parametric Bootstrap Solution to the MANOVA Under Heteroscedasticity." *Journal of Statistical Computation and Simulation* 80, no. 8: 873–887.

Landis, J. R., M. E. Miller, C. S. Davis, and G. G. Koch. 1988. "Some General Methods for the Analysis of Categorical Data in Longitudinal Studies." *Statistics in Medicine* 7, no. 1-2: 109–137.

Lawley, D. 1938. "A Generalization of Fisher's z Test." *Biometrika* 30, no. 1/2: 180–187.

Lekberg, Y., J. D. Bever, R. A. Bunn, et al. 2018. "Relative Importance of Competition and Plant–Soil Feedback, Their Synergy, Context Dependency and Implications for Coexistence." *Ecology Letters* 21, no. 8: 1268–1281.

Lin, D. 1997. "Non-Parametric Inference for Cumulative Incidence Functions in Competing Risks Studies." *Statistics in Medicine* 16, no. 8: 901–910.

Liu, R. Y. 1988. "Bootstrap Procedures Under Some Non-IID Models." *Annals of Statistics* 16, no. 4: 1696–1708.

Mammen, E. 1993. "Bootstrap and Wild Bootstrap for High Dimensional Linear Models." *Annals of Statistics* 21, no. 1: 255–285.

Martinussen, T., and T. H. Scheike. 2007. *Dynamic Regression Models for Survival Data*. New York: Springer Science & Business Media.

Micceri, T. 1989. "The Unicorn, the Normal Curve, and Other Improbable Creatures." *Psychological Bulletin* 105, no. 1: 156.

Molenberghs, G., and M. Kenward. 2007. *Missing Data in Clinical Studies*. Vol 61. Chichester, United Kingdom: John Wiley & Sons.

Molenberghs, G., M. G. Kenward, and E. Lesaffre. 1997. "The Analysis of Longitudinal Ordinal Data With Nonrandom Drop-Out." *Biometrika* 84, no. 1: 33–44.

Molenberghs, G., and E. Lesaffre. 1994. "Marginal Modeling of Correlated Ordinal Data Using a Multivariate Plackett Distribution." *Journal of the American Statistical Association* 89, no. 426: 633–644.

Molenberghs, G., and G. Verbeke. 2004. "Meaningful Statistical Model Formulations for Repeated Measures." *Statistica Sinica* 14, no. 3: 989–1020.

Munzel, U., and E. Brunner. 2000. "Nonparametric Methods in Multivariate Factorial Designs." *Journal of Statistical Planning and Inference* 88, no. 1: 117–132.

Nesnow, S., M. J. Mass, J. A. Ross, et al. 1998. "Lung Tumorigenic Interactions in Strain A/J Mice of Five Environmental Polycyclic Aromatic Hydrocarbons." *Environmental Health Perspectives* 106, no. suppl 6: 1337–1346.

Noguchi, K., Y. R. Gel, E. Brunner, and F. Konietschke. 2012. "nparLD: An R Software Package for the Nonparametric Analysis of Longitudinal Data in Factorial Experiments." *Journal of Statistical Software* 50, no. 12: 1–23.

Pan, R., T. Yang, J. Cao, K. Lu, and Z. Zhang. 2015. "Missing Data Imputation by K Nearest Neighbours Based on Grey Relational Structure and Mutual Information." *Applied Intelligence* 43, no. 3: 614–632.

Pauly, M., E. Brunner, and F. Konietschke. 2015. "Asymptotic Permutation Tests in General Factorial Designs." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 77, no. 2: 461–473.

Pauly, M., D. Ellenberger, and E. Brunner. 2015. "Analysis of High-Dimensional One Group Repeated Measures Designs." *Statistics* 49, no. 6: 1243–1261.

Pesarin, F., and L. Salmaso. 2012. "A Review and Some New Results on Permutation Testing for Multivariate Problems." *Statistics and Computing* 22, no. 2: 639–646.

Qian, H.-R., and S. Huang. 2005. "Comparison of False Discovery Rate Methods in Identifying Genes With Differential Expression." *Genomics* 86, no. 4: 495–503.

R Core Team. 2013. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rubarth, K., M. Pauly, and F. Konietschke. 2022. "Ranking Procedures for Repeated Measures Designs With Missing Data: Estimation, Testing and Asymptotic Theory." *Statistical Methods in Medical Research* 31, no. 1: 105–118.

Rubarth, K., P. Sattler, H. G. Zimmermann, and F. Konietschke. 2022. "Estimation and Testing of Wilcoxon–Mann–Whitney Effects in Factorial Clustered Data Designs." *Symmetry* 14, no. 2. https://www.mdpi.com/about/announcements/784.

Santos, M. S., R. C. Pereira, A. F. Costa, J. P. Soares, J. Santos, and P. H. Abreu. 2019. "Generating Synthetic Missing Data: A Review by Missing Mechanism." *IEEE Access* 7: 11651–11667.

Sattler, P. 2021. "A Comprehensive Treatment of Quadratic-Form-Based Inference in Repeated Measures Designs Under Diverse Asymptotics." *Electronic Journal of Statistics* 15, no. 1: 3611–3634.

Sattler, P., and M. Pauly. 2018. "Inference for High-Dimensional Split-Plot-Designs: A Unified Approach for Small to Large Numbers of Factor Levels." *Electronic Journal of Statistics* 12, no. 2: 2743–2805.

Smaga, Ł. 2017. "Bootstrap Methods for Multivariate Hypothesis Testing." *Communications in Statistics-Simulation and Computation* 46, no. 10: 7654–7667.

Traux, C., and R. R. Carkhuff. 1965. "Personality Change in Hospitalized Mental Patients During Group Psychotherapy as a Function of the Use of Alternate Sessions and Vicarious Therapy Pretraining." *Journal of Clinical Psychology* 21: 225–228.

Umlauft, M., M. Placzek, F. Konietschke, and M. Pauly. 2019. "Wild Bootstrapping Rank-Based Procedures: Multiple Testing in Nonparametric Factorial Repeated Measures Designs." *Journal of Multivariate Analysis* 171: 176–192.

Vallejo, G., M. Fernández, and P. E. Livacic-Rojas. 2010. "Analysis of Unbalanced Factorial Designs With Heteroscedastic Data." *Journal of Statistical Computation and Simulation* 80, no. 1: 75–88.

Van Steen, K., G. Molenberghs, G. Verbeke, and H. Thijs. 2001. "A Local Influence Approach to Sensitivity Analysis of Incomplete Longitudinal Ordinal Data." *Statistical Modelling* 1, no. 2: 125–142.

Wildsmith, S., G. Archer, A. Winkley, P. Lane, and P. Bugelski. 2001. "Maximization of Signal Derived From cDNA Microarrays." *Biotechniques* 30, no. 1: 202–208.

Xu, J., and X. Cui. 2008. "Robustified MANOVA With Applications in Detecting Differentially Expressed Genes From Oligonucleotide Arrays." *Bioinformatics* 24, no. 8: 1056–1062.

Xu, L.-W., F.-Q. Yang, A. Abula, and S. Qin. 2013. "A Parametric Bootstrap Approach for Two-Way ANOVA in Presence of Possible Interactions With Unequal Variances." *Journal of Multivariate Analysis* 115: 172–180.

Yan, J. 2007. "Enjoy the Joy of Copulas: With a Package Copula." *Journal of Statistical Software* 21, no. 4: 1–21.

Zhu, B., C. He, and P. Liatsis. 2012. "A Robust Missing Value Imputation Method for Noisy Data." *Applied Intelligence* 36, no. 1: 61–74.

**Supporting Information**

Additional supporting information can be found online in the Supporting Information section.