Received: 16 February 2020 | Revised: 11 December 2020 | Accepted: 20 January 2021

DOI: 10.1002/bimj.202000051

Biometrical Journal

RESEARCH PAPER

Asymptotic-based bootstrap approach for matched pairs with missingness in a single arm

Lubna Amro | **Markus Pauly** | **Burim Ramosaj**

Mathematical Statistics and Applications in Industry, Faculty of Statistics, Technical University of Dortmund, Dortmund, Germany

Correspondence
Lubna Amro, Mathematical Statistics and Applications in Industry, Faculty of Statistics, Technical University of Dortmund, 44227 Dortmund, Germany.
Email: lubna.amro@tu-dortmund.de

Funding information
Deutsche Forschungsgemeinschaft; Deutscher Akademischer Austauschdienst

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 could fully be reproduced.

Abstract

The issue of missing values is an arising difficulty when dealing with paired data. Several test procedures are developed in the literature to tackle this problem. Some of them are even robust under deviations and control type-I error quite accurately. However, most of these methods are not applicable when missing values are present only in a single arm. For this case, we provide asymptotic correct resampling tests that are robust under heteroskedasticity and skewed distributions. The tests are based on a meaningful restructuring of all observed information in quadratic form–type test statistics. An extensive simulation study is conducted exemplifying the tests for finite sample sizes under different missingness mechanisms. In addition, illustrative data examples based on real life studies are analyzed.

KEYWORDS
matched pairs, missing values, parametric bootstrap, quadratic forms

1 | INTRODUCTION

Conducting statistical tests on units measured repeatedly requires the consideration of the dependence structure of the resulting random vector. The simplest design is the matched pairs model, where units are measured at two endpoints of the same subject. This design has experienced a large field of application, including industrial and life sciences. In Biomedicine for example, several studies have been focused on identifying genes for up- or downregulated effects in head and neck squamous, prostate, lung, or breast cell carcinoma (Kuriakose et al., 2004; Lapointe et al., 2004; Feng et al., 2008). In common statistical analysis, testing the equality of means in matched pairs design is conducted using the paired t-test. Even for nonnormal data, the procedure is asymptotically exact, that is, for sufficiently large samples, the test procedure is correctly reflecting type-I error. However, first limitation of the paired t-test arises when data are only partially observed. Deleting observations with missing values is a suboptimal solution, since variance or mean estimation based only on complete case analysis can be biased leading to incorrect statistical inference. This is especially the case when complete samples are small.

To tackle this issue, a simple approach is to impute missing values singly (or multiply) and to carry out statistical tests as if there were no missing values so far (Schafer, 1999; Rubin, 2004; Sterne et al., 2009). However, although leading to good imputation error (Stekhoven & Bühlmann, 2011; Waljee et al., 2013; Ramosaj & Pauly, 2019), such approaches may lead to inflated type-I error rate or remarkably low power in small to moderate sample sizes (Van Buuren, 2018; Ramosaj et al., 2020). Therefore, we do not follow this approach here.

Differing to imputation, several test procedures that (only) use all observed information in the matched pairs design have been proposed in the literature (Mehta & Gurland, 1969; Lin, 1973; Morrison, 1973; Lin & Stivers, 1974; Little, 1976; Ekbohm, 1976; Bhoj, 1978; Looney & Jones, 2003; Kim et al., 2004; Xu & Harrar, 2012; Fuchs et al., 2017; Uddin & Hasan, 2017). These tests, however, rely on specific model assumptions such as symmetry or even bivariate normality, which are hard to verify in practice. Moreover, these procedures are usually nonrobust to deviations and may result in inaccurate decisions caused by possibly conservative or inflated type-I error rates (Samawi & Vogel, 2014; Amro & Pauly, 2017; Amro et al., 2019; Qi et al., 2019; Harrar et al., 2020).

To overcome these problems, the typical recommendation is to use the method based on combining separate results of adequate test statistics for the underlying paired and unpaired portions of the data using either weighted test statistics (Samawi & Vogel, 2014; Amro & Pauly, 2017; Martínez-Camblor et al., 2013), a multiplication combination test (Amro et al., 2019), or combined p-values (Rempala & Looney, 2006; Samawi et al., 2011; Yu et al., 2012; Kuan & Huang, 2013). However, all these methods are only applicable for matched pairs with missingness in both arms. This is due to their tests construction. Since, they are based upon combining the results of two independent tests for the related paired and unpaired two-sample problem. As independence of these two tests is required, a direct adjustment to handle data with missingness in one arm only is not possible. Thus, these methods cannot be used to analyze data on pathological stage I breast cancer patients from the Cancer Genome Atlas (TCGA) project. This data set consists of observations from 90 patients of which 74 had entries in one component of it, only 16 were complete, see Section 7.1 for details. The question is now how to analyze such data?

In contrast to the above methods, barely any work can be found that is potentially applicable in this special missing pattern, requires no parametric assumptions and also leads to valid inferences in case of heteroskedasticity or skewed distributions. One exception is given by the recent proposals of Qi et al. (2019) who recommended a so-called nonparametric combination test (NCT) and nonparametric p-value pooling methods (NPM). The NCT is based on merging the results from *Sign test* and *Wilcoxon Mann–Whitney test* while the NPM are based on combining p-values of the *Wilcoxon signed-rank test* and *Mann–Whitney test*. In situations where these two nonparametric procedures show their efficiency, their proposed combination is indeed tempting. However, neither the *Sign test* is known to be very powerful for metric data nor is the *Mann–Whitney test* known for being robust against heteroskedasticity. In fact, our simulation studies demonstrate that the NCT and Fisher's pooling method (FPM) as an NPM inherit these unsatisfying properties to some extent: under heteroskedasticity and/or skewed distributions, the NCT and FPM tend to not maintain the pre-assigned type-I error level. The degree of variance heterogeneity, skewness, and sample sizes can all affect the type-I error rate control level. An example of the type-I error control of NCT and FPM when heteroskedasticity coincides with a skewed error distribution is displayed in Figure 1. It reveals that, under heteroskedasticity and an exponential distribution, the NCT and FPM type-I error rate functions become surprisingly analogous to the power function where the type-I error rate increases dramatically with an increase in sample sizes.

The aim of this paper is therefore bilateral: First, we aim to provide a statistical test that is capable of treating single-arm missing values in matched pairs which drop the common assumptions such as homoskedasticity and normality, while not losing (partial) information. Second, it should be able to satisfactorily control type-I error while maintaining good power properties. To this end, we propose three different test statistics, analyze their asymptotic behaviors under the null hypothesis and equip them with an asymptotically correct parametric bootstrap procedure for calculating critical values. In doing so, we structured the paper by first introducing the statistical model and the hypothesis of interest. In Section 3, we provide different test statistics of quadratic form–type that either converge to a χ^2 or a weighted χ^2-distribution. Proofs presenting theoretical guarantees of the proposed methods are delivered in the supplement. In Section 4, we introduce a parametric bootstrap technique to calculate critical values and prove its theoretical correctness. Section 5 is devoted to already existing methods for statistical inference in matched pairs with single-arm missingness while in Sections 6 and 7, novel and existing methods are compared based on an extensive simulation study and three real life data examples. The supplement contains additional theoretical details. For notational purposes, we state vectors or matrices in bold and scalars in usual form.

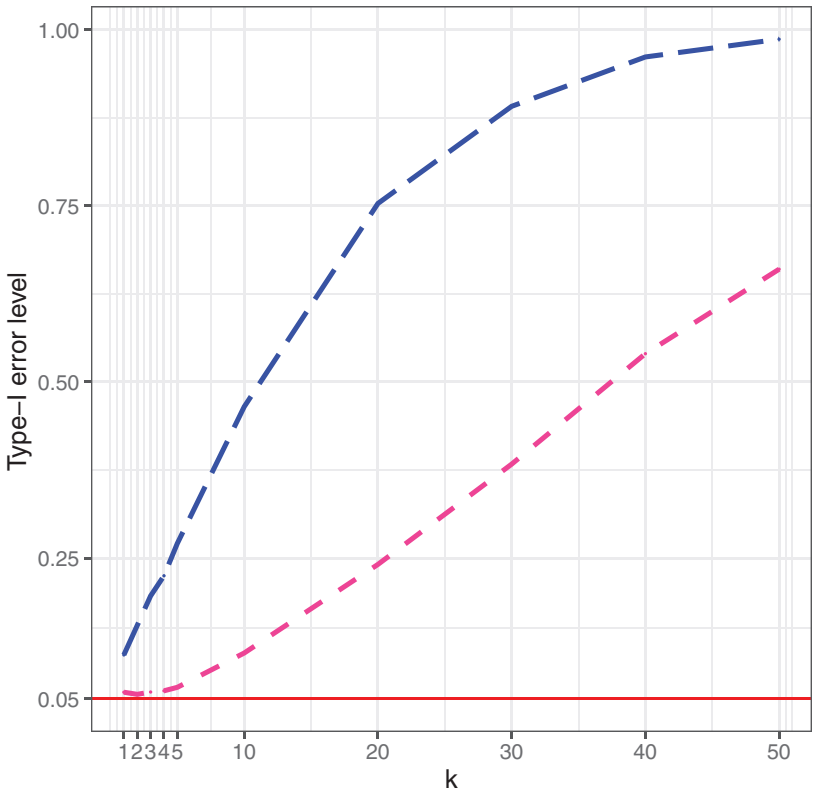

FIGURE 1 Type-I error simulation results ($\alpha = .05$) of the nonparametric combination test T_N (— —) and the Fisher's p-value pooling method T_F (– – –) for exponential distribution under correlation factor ($\rho = .7$) and a heteroskedastic setup with variances 1 and 2, respectively, for increasing sample sizes $k \cdot (n_c, n_u) = (k \cdot 10, k \cdot 30)$ under the MCAR framework

2 | STATISTICAL MODEL AND HYPOTHESES

We consider matched pairs given by a sample $\mathcal{D}_n := \{\mathbf{X}_1, \dots, \mathbf{X}_n\}$, where $\mathbf{X}_j = [X_{1j}, X_{2j}]^\top \in \mathbb{R}^2$ are i.i.d. random vectors with mean vector $\mathbb{E}[\mathbf{X}_1] = \boldsymbol{\mu} = [\mu_1, \mu_2]^\top \in \mathbb{R}^2$ and an arbitrary covariance matrix $0 < \boldsymbol{\Gamma} = \begin{bmatrix} \tilde{\sigma}_1^2 & \rho\tilde{\sigma}_1\tilde{\sigma}_2 \\ \rho\tilde{\sigma}_1\tilde{\sigma}_2 & \tilde{\sigma}_2^2 \end{bmatrix} \in \mathbb{R}^{2\times 2}$, where $\tilde{\sigma}_1^2 = \text{var}(X_{11})$, $\tilde{\sigma}_2^2 = \text{var}(X_{21})$ and $\rho = \text{corr}(X_{11}, X_{21})$. To incorporate missingness in one arm (says, the second) only denote with $R_{2j} \in \{0, 1\}$, $j = 1, \dots, n$ the vector whose jth component indicates whether X_{2j} is observed ($R_{2j} = 1$) or missing ($R_{2j} = 0$) for $j = 1, \dots, n$. Define the composition $*$ by $a * 1 = a$ and $a * 0 = ---$, for all $a \in \mathbb{R}$, then in practice, one observes $\mathbf{X}^{(o)} := \{\mathbf{X}_j * \mathbf{R}_j\}_{j=1}^n$ where $\mathbf{R}_j = [1, R_{2j}]^\top \in \mathbb{R}^2$, $j = 1, \dots, n$, and a "$---$" entry is interpreted as missing. Hence, our framework has the following form:

$$\underbrace{\begin{bmatrix} X_{11}^{(c)} \\ X_{21}^{(c)} \end{bmatrix}, \dots, \begin{bmatrix} X_{1n_c}^{(c)} \\ X_{2n_c}^{(c)} \end{bmatrix}}_{\mathbf{X}^{(c)}}, \underbrace{\begin{bmatrix} X_{11}^{(i)} \\ --- \end{bmatrix}, \dots, \begin{bmatrix} X_{1n_u}^{(i)} \\ --- \end{bmatrix}}_{\mathbf{X}^{(i)}}. \tag{1}$$

Rubin defines the missing mechanism through a parametric distributional model on $\mathbf{R} = \{\mathbf{R}_j\}_{j=1}^n$ and classifies their presence through Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing not at Random (MNAR) schemes (Rubin, 2004). In our work, we first assume an MCAR mechanism, in that $\boldsymbol{X}^{(c)}$ is independent of $\boldsymbol{X}^{(i)}$. However, we will also study MAR mechanisms in simulations and relate to the supplement for the explicit definition of the missing mechanisms. For notational purposes, let I_{n_c} denote the index set of $|I_{n_c}| = n_c$ complete pairs, that is, $\mathbf{R}_j = [1, 1]^\top$ for all $j \in I_{n_c}$. Similarly, I_{n_u} is the index set of observations with second component missing ($\mathbf{R}_j = [1, 0]^\top$, $j \in I_{n_u}$) and $|I_{n_u}| = n_u$. Thus, there are in total $N = 2n_c + n_u$ observations from $n = n_c + n_u$ subjects.

In this framework, we would like to use all the available data to test the null hypothesis $H_0 : \{\mu_1 = \mu_2\}$ of equal means against the alternative $H_1 : \{\mu_1 \neq \mu_2\}$.

To construct our test statistics, we first fix estimators of the population means μ_1, and μ_2. For estimating μ_1, we consider two estimators; the sample mean of the first components of the completed data set $\bar{X}_{1\cdot}^{(c)} = \frac{1}{n_c}\sum_{i=1}^{n_c} X_{1i}^{(c)}$, and the sample

mean of the first components of the unpaired data $\bar{X}_{1.}^{(i)} = \frac{1}{n_u}\sum_{j=1}^{n_u} X_{1j}^{(i)}$. For estimating the population mean μ_2, we use the sample mean of the second components of the complete data $\bar{X}_{2.}^{(c)} = \frac{1}{n_c}\sum_{i=1}^{n_c} X_{2i}^{(c)}$. Next, we define the normalized vector $\boldsymbol{Z}_n$ that aggregates the difference between the mean values $\boldsymbol{\mu} = [\mu_1, \mu_2]^\top$ and their empirical estimators $[\bar{X}_{1.}^{(c)}, \bar{X}_{2.}^{(c)}, \bar{X}_{1.}^{(i)}]^\top$

$$\boldsymbol{Z}_n = \sqrt{n}[\bar{X}_{1.}^{(c)} - \mu_1, \bar{X}_{2.}^{(c)} - \mu_2, \bar{X}_{1.}^{(i)} - \mu_1]^\top \tag{2}$$

and take their correlations into account in the covariance matrix

$$\boldsymbol{\Sigma}_n := \mathrm{cov}(\boldsymbol{Z}_n) = \begin{bmatrix} (n/n_c)\sigma_1^2 & (n/n_c)\rho\sigma_1\sigma_2 & 0 \\ (n/n_c)\rho\sigma_1\sigma_2 & (n/n_c)\sigma_2^2 & 0 \\ 0 & 0 & (n/n_u)\sigma_1^2 \end{bmatrix},$$

where $\sigma_1^2 = \mathrm{var}(X_{11}^{(c)}) = \mathrm{var}(X_{11}^{(i)})$, $\sigma_2^2 = \mathrm{var}(X_{21}^{(c)})$, and $\rho = \mathrm{corr}(X_{11}^{(c)}, X_{21}^{(c)})$.

To test the null hypothesis $H_0 : \{\mu_1 - \mu_2 = 0\}$, we define the two estimators $\bar{X}_{1.}^{(c)} - \bar{X}_{2.}^{(c)}$ and $\bar{X}_{1.}^{(i)} - \bar{X}_{2.}^{(c)}$ for $\mu_1 - \mu_2$. Their joined asymptotic behavior under the null hypothesis H_0 is studied below.

Proposition 1. *Set $f_{\boldsymbol{A}}(\mathbf{x}) = \boldsymbol{A}\mathbf{x}$, for the matrix $\boldsymbol{A} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \in \mathbb{R}^{2X3}$. Then, under the null hypothesis H_0 and the condition that $\frac{n_c}{n_c+n_u} \to \kappa_1 \in (0,1)$ and $\frac{n_u}{n_c+n_u} \to \kappa_2 = (1-\kappa_1) \in (0,1)$ as $n \to \infty$, the composite statistic*

$$f_{\boldsymbol{A}} \circ \boldsymbol{Z}_n = \mathbf{A}\mathbf{Z}_n = \sqrt{n}[\bar{X}_{1.}^{(c)} - \bar{X}_{2.}^{(c)}, \bar{X}_{1.}^{(i)} - \bar{X}_{2.}^{(c)}]^\top \tag{3}$$

is asymptotically $N_2(\mathbf{0}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)$ distributed as $n \to \infty$.

$$\textit{Here, } \boldsymbol{\Sigma} = \lim_{n\to\infty} \boldsymbol{\Sigma}_n = \begin{bmatrix} \kappa_1^{-1}\sigma_1^2 & \kappa_1^{-1}\rho\sigma_1\sigma_2 & 0 \\ \kappa_1^{-1}\rho\sigma_1\sigma_2 & \kappa_1^{-1}\sigma_2^2 & 0 \\ 0 & 0 & \kappa_2^{-1}\sigma_1^2 \end{bmatrix}. \tag{4}$$

3 | STATISTICS AND ASYMPTOTICS

In this section, we propose three different quadratic forms for testing H_0: a Wald-type statistic (WTS), an ANOVA-L2-type statistic (ATS), and a modified ANOVA-type statistic (MATS). To introduce the WTS, denote by $\mathbf{B}^+$ the Moore–Penrose inverse of a matrix $\mathbf{B}$. Then, the WTS is given by

$$T_W = (\boldsymbol{A}\boldsymbol{Z}_n)^\top (\boldsymbol{A}\widehat{\boldsymbol{\Sigma}}_n\boldsymbol{A}^\top)^+ (\boldsymbol{A}\boldsymbol{Z}_n), \tag{5}$$

where $\widehat{\boldsymbol{\Sigma}}_n$ is the plug-in sample estimator for $\boldsymbol{\Sigma}$ given in (4), see the supplement for its explicit form. Thanks to the introduced studentization by $(\boldsymbol{A}\widehat{\boldsymbol{\Sigma}}_n\boldsymbol{A}^\top)^+$, the WTS is asymptotically χ_2^2-distributed under the null hypothesis as long as $\boldsymbol{\Sigma} > 0$ as proved in the supplement.

Similar WTS versions are also studied in the context of heteroskedastic ANOVA or MANOVA (Krishnamoorthy & Lu, 2010; Xu et al., 2013; Konietschke et al., 2015; Friedrich & Pauly, 2018). From these settings, it is known that the convergence to its limiting χ^2-distribution is rather slow and large sample sizes are required to obtain adequate results (Vallejo et al., 2010; Konietschke et al., 2015; Smaga, 2017), which leads to several refinements regarding bootstrapping for the calculations of critical values (see Section 4) or other structures of test statistics. In particular Brunner (2001) proposed an alternative quadratic form by deleting the estimated covariance matrix $\widehat{\boldsymbol{\Sigma}}_n$ involved in the computation of the WTS. Here, we erase the Moore–Penrose inverse term from the WTS resulting in the following ATS:

$$T_A = \frac{1}{tr(\mathbf{A}\widehat{\boldsymbol{\Sigma}}_n\mathbf{A}^\top)} (\mathbf{A}\boldsymbol{Z}_n)^\top (\mathbf{A}\boldsymbol{Z}_n). \tag{6}$$

The ATS has the advantage of being applicable in case of singular covariance matrices ($|\widehat{\boldsymbol{\Sigma}}_n| = 0$). However, it has the drawback of not being asymptotically distribution-free under the null hypothesis, see the supplement for details.

Another possible test statistic would be the MATS that was developed by Friedrich & Pauly (2018) for MANOVA models. The authors could provide preferable simulation results regarding its power behavior and type-I error control while delivering theoretical guarantees for its validity. Hence, we consider a MATS (with a slight modification) in our design, too. Here, it is given by

$$T_M = (\boldsymbol{AZ}_n)^\top \widehat{\boldsymbol{D}}_n (\boldsymbol{AZ}_n), \tag{7}$$

where $\widehat{\boldsymbol{D}}_n = \mathrm{diag}((\mathbf{A}\widehat{\boldsymbol{\Sigma}}_n\mathbf{A}^\top)^+_{ii})$.

Similar to the ATS, the MATS is also not distribution-free under H_0, see the supplement for the explicit form of its limiting distribution. Thus, we cannot directly calculate critical values for T_A and T_M, respectively. In addition, the χ^2_2-approximation to T_W is rather slow. To this end, we develop adequate and asymptotically correct testing procedures based on bootstrap versions of T_W, T_A, and T_M in the subsequent section.

4 | PARAMETRIC BOOTSTRAPPING

To estimate critical values, we apply an asymptotic model-based bootstrap approach which has, for example, been applied in the context of (M)ANOVA factorial designs (Konietschke et al., 2015; Friedrich & Pauly, 2018). To this end, we first generate parametric bootstrap variables as

$$\mathbf{X}^*_j = \begin{bmatrix} X^*_{1j} \\ X^*_{2j} \end{bmatrix} \overset{\text{i.i.d}}{\sim} N(0, \widehat{\boldsymbol{\Gamma}}), j = 1, \dots, n. \tag{8}$$

Here, $\widehat{\boldsymbol{\Gamma}} = \begin{bmatrix} \widehat{\sigma}^2_1 & \widehat{\rho}\widehat{\sigma}_1\widehat{\sigma}_2 \\ \widehat{\rho}\widehat{\sigma}_1\widehat{\sigma}_2 & \widehat{\sigma}^2_2 \end{bmatrix}$ is the empirical covariance matrix estimator, that is, $\widehat{\sigma}^2_i$ denotes the sample variance calculated on all observations in component i and $\widehat{\rho}$ is the sample correlation obtained from $\boldsymbol{X}^{(c)}$. The idea is to reflect the original covariance structure to obtain more accurate finite sample approximation. Next, we generate missing values under the MCAR scheme by randomly inserting them to the second component of the bivariate vector $\mathbf{X}^*_j$ until a fixed amount of missing values of size n_u is achieved. This results into the following bootstrapped data set:

$$\underbrace{\begin{bmatrix} X^{*(c)}_{11} \\ X^{*(c)}_{21} \end{bmatrix}, \dots, \begin{bmatrix} X^{*(c)}_{1n_c} \\ X^{*(c)}_{2n_c} \end{bmatrix}}_{\mathbf{X}^{*(c)}}, \underbrace{\begin{bmatrix} X^{*(i)}_{11} \\ --- \end{bmatrix}, \dots, \begin{bmatrix} X^{*(i)}_{1n_u} \\ --- \end{bmatrix}}_{\mathbf{X}^{*(i)}} \tag{9}$$

and the combined vector $(f \circ \boldsymbol{Z}_n)^* = \boldsymbol{AZ}^*_n = \sqrt{n}(\bar{X}^{*(c)}_{1.} - \bar{X}^{*(c)}_{2.}, \bar{X}^{*(i)}_{1.} - \bar{X}^{*(c)}_{2.})$. 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:

$$T^*_W = (\boldsymbol{AZ}^*_n)^\top (\mathbf{A}\widehat{\boldsymbol{\Sigma}}^*_n\mathbf{A}^T)^+ (\boldsymbol{AZ}^*_n), \tag{10}$$

$$T^*_A = \frac{1}{tr(\mathbf{A}\widehat{\boldsymbol{\Sigma}}^*_n\mathbf{A}^\top)} (\boldsymbol{AZ}^*_n)^\top (\boldsymbol{AZ}^*_n), \tag{11}$$

$$T^*_M = (\boldsymbol{AZ}^*_n)^\top \widehat{\boldsymbol{D}}^*_n (\boldsymbol{AZ}^*_n), \tag{12}$$

where $\widehat{\boldsymbol{\Sigma}}^*_n = \widehat{\boldsymbol{\Sigma}}_n(\mathbf{X}^{*(c)}, \mathbf{X}^{*(i)})$ and $\widehat{\boldsymbol{D}}^*_n = \mathrm{diag}((\mathbf{A}\widehat{\boldsymbol{\Sigma}}^*_n\mathbf{A}^\top)^+_{ii})$.

It is proven in the supplement that all three bootstrapped test statistics approximate the null distribution of the respective test statistic.

To analyze their finite sample performance, we below conduct extensive simulations (Section 6). Before that, we will first discuss other possible candidates from the literature that should or should not be included in our simulation study.

5 | COMPARISON WITH EXISTING MODELS

We briefly review the existing literature on methods that can deal with the case of matched pairs with missing values in one arm only. As outlined in the introduction, there only exists a few which we can summarize as follows:

(a) Simple methods such as: using the paired t-test while excluding the unpaired data OR using the independent t-test while ignoring the covariance structure of the data.
(b) Tests based on modified maximum likelihood estimators (Morrison, 1973; Ekbohm, 1976; Little, 1976).
(c) Tests based on simple mean difference estimators (Mehta & Gurland, 1969, 1973; Lin, 1973; Ekbohm, 1976).
(d) p-Values pooling methods (Qi et al., 2019).
(e) Weighted linear and nonlinear combination tests (Pesarin & Salmaso, 2010; Qi et al., 2019).

However, none of the methods is free from distributional assumptions and at the same time robust against deviations such as heteroskedasticity and skewed distributions. In particular, the recent paper by Qi et al. (2019) already included a simulation study to compare several of the tests mentioned in (a)–(e). As a conclusion, they recommended a so-called NCT and p-value pooling methods.

They investigated in their paper two ways of combining the p-values; a weighted inverse normal method proposed by Hartung (1999) and an FPM suggested by Brown (1975), Kost & McDermott (2002), and Hou (2005). Due to their quite similar behavior, we only include the FPM and the NCT into our simulation study. As additional competitor for these two and the bootstrap procedures proposed in Section 4, we choose the test of Little (1976). The latter assumes that the data follow a bivariate normal distribution and the test statistic is given by

$$T_L = \frac{\bar{X}_{1\cdot} - \bar{X}_{2\cdot}^{(c)} - \frac{\hat{\rho}\hat{\sigma}_1^{(c)}\hat{\sigma}_2}{\left(\hat{\sigma}_1^{(c)}\right)^2}\left(\bar{X}_{1\cdot} - \bar{X}_{1\cdot}^{(c)}\right)}{\hat{\sigma}_{\mathbf{L}}}, \tag{13}$$

where $\bar{X}_1 := 1/n(n_c\bar{X}_{1\cdot}^{(c)} + n_u\bar{X}_{1\cdot}^{(i)})$ and $\hat{\sigma}_1^{(c)}$ is the empirical standard deviation of $\{X_{11}^{(c)}, \ldots, X_{1n_c}^{(c)}\}$. Moreover, setting $\hat{\sigma}_{22\cdot1}^2 = \hat{\sigma}_2^2 - (\hat{\rho}\hat{\sigma}_1^{(c)}\hat{\sigma}_2/(\hat{\sigma}_1^{(c)})^2)$ and $\hat{\sigma}_{\mathbf{X}} = \hat{\sigma}_{22\cdot1}^2 + \frac{(\hat{\rho}\hat{\sigma}_1^{(c)}\hat{\sigma}_2)^2}{(\hat{\sigma}_1^{(c)})^4}\hat{\sigma}_1^4$, the denominator is given by Little (1976)

$$\hat{\sigma}_{\mathbf{L}}^2 = \frac{\hat{\sigma}_{\mathbf{X}}^2}{n} + \left(\frac{1}{n_c} - \frac{1}{n}\right)\frac{n_c - 2}{n_c - 3}\hat{\sigma}_{22\cdot1}^2 - \frac{2}{n}\frac{\hat{\rho}\hat{\sigma}_1^{(c)}\hat{\sigma}_2}{\left(\hat{\sigma}_1^{(c)}\right)^2}\hat{\sigma}_1^2 + \frac{\hat{\sigma}_1^2}{n}. \tag{14}$$

The exact distribution of T_L is rather complicated and Little suggests to approximate it by a t-reference distribution with $n_c - 1$ degrees of freedom, that is, the test is given by $\varphi_L := \mathbb{1}\{|T_L| > t_{n_c-1,1-\alpha/2}\}$ for some level $\alpha \in (0,1)$. To enhance its small sample properties (see the simulation results for φ_L given in the supplement for details), a parametric bootstrap version of the Little test is studied as well. Similar to φ_L, the resulting Little bootstrap test, $\varphi_L^* := \mathbb{1}\{|T_L| > c_L^*\}$ is asymptotically correct. Here, c_L^* denotes the conditional $(1-\alpha)$-quantile of the parametric bootstrap distribution of T_L.

In addition, the NCT proposed by Qi et al. (2019), is based upon a linear combination of the sign and the Wilcoxon Mann–Whitney test statistics:

$$T_N = T_s + T_m, \tag{15}$$

where $T_s = \frac{1}{n_c}\sum_{i=1}^{n_c}\phi(X_{1i}^{(c)}, X_{2i}^{(c)})$ and $T_m = \frac{1}{n_c n_u}\sum_{j=1}^{n_u}\sum_{k=1}^{n_c}\phi(X_{1j}^{(i)}, X_{2k}^{(c)})$ with $\phi(X_1, X_2) = \begin{cases} 1 & \text{if } X > Y, \\ 1/2 & \text{if } X = Y, \\ 0 & \text{otherwise.} \end{cases}$

It is proposed to approximate the null distribution of T_N by a normal distribution with mean 1 and variance estimated by $\widehat{\text{var}}(T_N) = \frac{1}{n_c} + \frac{n_c+n_u+1}{12n_cn_u} + \widehat{\text{cov}}(T_s, T_m)$, where
$\widehat{\text{cov}}(T_s, T_m) = \frac{1}{n_c^2 n_u} \sum_{i=1}^{n_c} \sum_{j=1}^{n_u} \mathbb{1}\{X_{1i}^{(c)} > X_{2i}^{(c)}, X_{1j}^{(i)} > X_{2j}^{(c)}\} - \frac{1}{n_c} T_s T_m$.

Moreover, the NPM proposed by Qi et al. (2019) based upon Fisher's pooling approach is based upon combining the dependent p-values of the Wilcoxon signed-rank test P_p and Mann–Whitney U test P_{up}. The test statistic is given by

$$T_F = -2\lambda_1 \log(P_p) - 2\lambda_2 \log(P_{up}), \tag{16}$$

where λ_1 and λ_2 are weights. It was shown that T_F follows asymptotically a scaled $c\chi_f^2$-distribution with $c = \frac{\text{var}(T_F)}{2E(T_F)}$ and $f = \frac{2[E(T_F)]^2}{\text{var}(T_F)}$. Moreover, the mean and variance of T_F are $E(T_F) = 2(\lambda_1 + \lambda_2)$, $\text{var}(T_F) = 4(\lambda_1^2 + \lambda_2^2) + 2\lambda_1\lambda_2\eta$, and $\eta = Cov(-2\log(P_p), -2\log(P_{up}))$ Qi et al. (2019). suggested to estimate η by nonparametric bootstrapping to obtain estimates $\hat{c}$ and $\hat{f}$ for c and f, respectively. Therefore, the null distribution of T_F can be asymptotically approximated by $\hat{c}\chi_{\hat{f}}^2$. In previous simulation studies by Qi et al. (2019), the considered choices of the weights λ_1 and λ_2 had almost invariant impact on the behavior of FPM. Similar to Qi et al. (2019), we therefore consider the following weights: $\lambda_1 = \sqrt{2n_c}$ and $\lambda_2 = \sqrt{n_c + n_u}$.

Inspired by Pesarin & Salmaso (2010), we also consider a nonparametric combination (NPC) of two dependent permutation tests. Their methodology is based upon properly breaking down a testing problem into a set of simpler subproblems. Then, each subproblem is provided with a proper permutation test, and jointly analyzed to maintain any underlying dependencies. Fitting this approach to our model, we choose a permutation paired t-test (Janssen, 1999; Konietschke & Pauly, 2014) that is computed upon the complete pairs $\mathbf{X}^{(c)}$ only and a permutation Welch-test (Janssen, 1997; Chung et al., 2013; Pauly et al., 2015) that is based upon $X_{1j}^{(i)}$, and $X_{2k}^{(c)}$. The global p-value is then obtained through combining the partial p-values of the above tests using Fisher's combining function. We denote this testing procedure by T_P. For more details about the NPC procedure and related R codes, we refer to the monographs of Pesarin (2001) and Pesarin & Salmaso (2010).

Finally, we also consider the most simple solution: the paired t-test T_t, calculated on the complete cases $\mathbf{X}^{(c)}$ only. We compare the finite sample performance of all these methods and the three new bootstrap approaches from Section 4 in the sequel. To judge the performance of all methods, a parametric bootstrap version of the paired t-test handling full data before introducing missingness has been included in all tables. The corresponding procedure is denoted by F.

6 | SIMULATION STUDY

In this section, we investigate the finite sample behavior of the methods described in Sections 4 and 5 in extensive simulations. 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.

Small- to moderate-sized paired data samples were generated from the model
$\mathbf{X}_j = \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{\varepsilon}_j + \boldsymbol{\mu}, \quad j = 1, \ldots, n,$
where $\boldsymbol{\varepsilon}_j = [\varepsilon_{1j}, \varepsilon_{2j}]^\top$ is an i.i.d. bivariate random vector with mutually independent components and $E(\boldsymbol{\varepsilon}_1) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\varepsilon}_1) = I_2$.

Different choices of symmetric as well as skewed residuals are considered such as standardized normal, exponential, Laplace, and the χ^2-distribution with $df = 30$ degrees of freedom. For the covariance matrix $\boldsymbol{\Sigma}$, we considered the choices
$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ and $\boldsymbol{\Sigma}_2 = \begin{bmatrix} 1 & \sqrt{2}\rho \\ \sqrt{2}\rho & 2 \end{bmatrix}$
with varying correlation factor $\rho \in (-1, 1)$, representing a homoskedastic and a heteroskedastic covariance setting, respectively. The sample sizes were chosen as $(n_c, n_u) \in \{(10, 10), (30, 10), (10, 30)\}$ under an MCAR mechanism and $n \in \{10, 20, 30\}$ under an MAR mechanism.

For each scenario, we generated missings as described below: For the *MCAR mechanism*, missing values are inserted randomly to the second component of the bivariate vector $\mathbf{X}_j$ until a fixed amount of missing values of size n_u for the second component is achieved.

For the *MAR mechanism*, the probability of being missing on the second component of $\mathbf{X}_j$ is based on the corresponding value on the first component in the following way: first, we divide $\mathbf{X}$ into three groups based on their first component values corresponding to a 2σ-rule: the first group is given by $\{\boldsymbol{X}_j = (X_{1j}, X_{2j}) : X_{1j} \in (-\infty, -2\sigma_1), j = 1, .., n\}$, the second by $\{\boldsymbol{X}_j : X_{1j} \in (-2\sigma_1, 2\sigma_1), j = 1, .., n\}$, and the last group by $\{\boldsymbol{X}_j : X_{1j} \in (2\sigma_1, \infty), j = 1, .., n\}$, where σ_1 is the variance of the first component. Then, we randomly insert missing values on the second component based on the following missing percentages: 15% for group one and three and 30% for the second group.

In order to assess the power of all methods, we set $\boldsymbol{\mu} = [\delta, 0]^\top$ with shift parameter $\delta \in \{0, 1/2, 1\}$. All simulations were operated by means of the statistical computing environment R based on $n_{sim} = 10,000$ Monte-Carlo runs and $B = 999$ bootstrap runs (in case of the three bootstrapped methods based upon T^*_W, T^*_A, and T^*_M and the bootstrapped version of Little's method T^*_L). The algorithm for the computation of the p-value of the parametric bootstrap tests is as follows:

1. For the given incomplete paired data, calculate the observed test statistic, say T.
2. Estimate the covariance matrix $\mathbf{\Gamma}$ by $\widehat{\mathbf{\Gamma}}$.
3. Generate a bootstrap sample $\mathbf{X}^*_j = (X^*_{1j}, X^*_{2j})$ from $N(\mathbf{0}, \widehat{\mathbf{\Gamma}})$, $j = 1, \dots, n$.
4. Insert missing values in an MCAR or MAR manner to the second component of the vector $\mathbf{X}^*_j$ resulting in $\mathbf{X}^{*(c)}_j$ and $\mathbf{X}^{*(i)}_k$ where $j = 1, \dots, n_c$, $k = 1, \dots, n_u$.
5. Calculate the value of the test statistic for the bootstrapped sample T^*.
6. Repeat the Steps 3 and 4 independently $B = 999$ times and collect the observed test statistic values in $T^*_b, b = 1, \dots .., B$.
7. Finally, estimate the bootstrap p-value as $P\text{-value} = \frac{\sum_{b=1}^{B} I(T^*_b >= T)}{B}$.

Now, the nonparametric bootstrap method that is used for estimating the covariance η of the Fisher's pooling method as suggested by Qi et al. (2019) is as follows:

1. Draw n_c times with replacement from the pairs $\mathbf{X}^{(c)}_j = (X^{(c)}_{1j}, X^{(c)}_{2j})$, $j = 1, \dots, n_c$, and calculate the p-value P^*_p.
2. Draw n_u times with replacement from $\mathbf{X}^{(i)}_k$, $k = 1, \dots, n_u$, and calculate the p-value P^*_{up}.
3. Replicate Step 1, $B = 999$ times and collect the observed p-values of the Wilcoxon signed-rank test (paired data) and Mann–Whitney U test (unpaired data) in P^*_{pb} and P^*_{ub}, respectively, $b = 1, \dots, B$.
4. Finally, estimate the parameter η needed for estimating the degrees of freedom as $\eta = \widehat{\text{cov}}(-2\log(\mathbf{P}^*_p), -2\log(\mathbf{P}^*_{up}))$, where $\mathbf{P}^*_p = \{P^*_{pb}, b = 1, \dots, B\}$ and $\mathbf{P}^*_{up} = \{P^*_{ub}, b = 1, \dots, B\}$.

Type-I Error Results. Simulation results of type-I error level of the studied procedures under the MCAR framework for different sample sizes and for homoskedastic as well as heteroskedastic settings are summarized in Tables 1, S.1, and S.2.

It can be readily seen that the suggested bootstrap approaches based upon T^*_W, T^*_A and T^*_M tend to result in quite accurate type-I error rate control under homoskedasticity as well as heteroskedasticity and over the whole range of correlation factors for most settings. Only in two cases; First, in case of the negative unbalanced sample size (10,30), particularly under heteroskedasticity, the bootstrapped MATS (T^*_M) is not recommended due to its liberal behavior. However, in this case, the other two suggested bootstrapped tests T^*_W, and T^*_A are controlling type-I error rate accurately. Secondly, in case of the skewed exponential distribution, the control is not adequate and a liberal behavior is observed. However, in this case, all the other chosen procedures also failed to control type-I error rate for the underlying sample sizes, which are indicated in bold red through all tables. Specifically, in the case of homoskedasticity, and a balanced sample size (10,10), our three suggested tests still result in accurate test decisions. For a positive balanced sample size (30,10), the bootstrapped ATS (T^*_A) still controls type-I error rate accurately under homoskedastic as well heteroskedastic settings. It has even the best control of type-I error rate under heteroskedasticity among all considered methods that are identified by bold entries in the table.

TABLE 1 Type-I error simulation results ($\alpha = .05$) of the tests for different distributions under varying correlation values (ρ) with sample sizes $(n_c, n_u) = (10, 10)$ and different covariance matrices Σ_1 and Σ_2 under the MCAR framework

		Σ_1									Σ_2								
			Parametric bootstrap				Alternatives					Parametric bootstrap				Alternatives			
Dist	ρ	F	T_W^*	T_A^*	T_M^*	T_L^*	T_t	T_N	T_F	T_P	F	T_W^*	T_A^*	T_M^*	T_L^*	T_t	T_N	T_F	T_P
Normal	−.9	5.3	5.3	5.2	5.4	**5.0**	4.8	6.7	4.3	7.5	5.1	**5.0**	5.3	5.6	4.9	4.7	7.1	4.8	8.3
	−.5	5.3	5.3	5.7	5.6	5.3	**5.1**	6.8	4.6	7.4	5.3	5.3	5.5	6	5.2	**5.0**	6.8	5.1	8
	−.1	5.3	4.6	4.9	4.8	4.8	**5.0**	6.5	4	6.3	5.4	4.6	5.5	5.2	**5.0**	5.1	7.2	4.6	7.7
	.1	4.9	4.8	5.4	5.1	**5.0**	4.8	6.4	4.3	6.6	**4.9**	**5.1**	5.5	5.4	**5.1**	**4.9**	7	**4.9**	7.7
	.5	5.3	5.4	**5.1**	**5.1**	4.3	5.3	6.2	4.3	5.9	5.2	5.4	5.6	5.5	**4.9**	5.2	7.1	4.7	7.3
	.9	5.3	5.2	**5.0**	4.3	4.5	5.4	5.7	4.2	**5.0**	5.2	5.1	**5.0**	5.9	3.2	5.4	6.9	4.5	7.1
Laplace	−.9	**4.9**	4.4	**4.9**	5.5	4.6	4.8	6.5	4.5	7.5	**4.9**	4.4	**5.1**	5.7	4.4	4.7	7.1	4.8	8
	−.5	5.1	4.4	5.2	5.1	**5.0**	4.5	6.6	4.3	7.3	5.1	4.4	5.1	5.4	**5.0**	4.6	7	4.7	7.9
	−.1	**4.9**	4.2	**4.9**	4.8	4.6	4.6	6.4	4.2	6.5	**5.0**	4.4	**5.0**	**5.0**	4.6	4.5	6.8	4.5	7.4
	.1	**4.8**	4.3	4.3	4.2	4.3	4.4	6.2	4	6.1	**4.9**	4.3	4.6	4.6	4.3	4.5	6.6	4.3	7
	.5	**5.1**	4.4	4.5	4.4	3.6	4.5	6.2	4	5.8	**4.9**	4.4	4.6	4.5	3.7	4.5	6.7	4.2	6.8
	.9	4.8	3.9	4.8	3.6	4.7	4.4	5.6	4	**4.9**	**4.8**	4.1	4.7	5.4	3.9	4.4	6.6	4.1	6.4
Exponential	−.9	**4.8**	4.7	4.4	5.6	4.5	4.2	6.5	4.3	6.8	**5.0**	4.6	5.2	6.8	5.3	4.7	8.7	4.1	8.2
	−.5	5.2	**5.1**	**4.9**	4.8	5.3	4.2	6.4	4.4	7.1	5.4	5.4	6.4	6.3	6.7	**5.0**	9.7	4.8	8.9
	−.1	5.3	5.3	**5.0**	4.6	5.8	4.4	6.6	4.3	6.9	5.6	6.1	6.6	6.1	7.1	**5.1**	10.1	4.8	8.7
	.1	**5.0**	**5.0**	4.4	4.1	5.9	4	6.6	4.1	6.2	5.6	6.1	6.8	5.9	7.6	5.2	10.5	**5.0**	8.7
	.5	**5.1**	5.8	4.5	4.2	6.5	4.2	6.2	4.7	6.1	5.9	7	6.9	6.5	7.7	5.7	10.7	**5.1**	8.8
	.9	**4.7**	5.8	4.4	3.6	7.3	4.1	5.5	4.4	4.4	7.5	8	**5.4**	8.5	7.8	8.8	12.1	5.6	9.7
Chi-square	−.9	**5.2**	5.4	5.6	5.8	**5.2**	**5.2**	6.9	**4.8**	7.8	5.4	5.5	5.8	6.1	**5.0**	5.3	7.6	5.2	8.7
	−.5	5.4	**5.0**	5.2	5.1	**5.0**	4.9	6.5	4.1	6.9	5.4	**5.0**	5.3	5.6	**5.0**	4.8	7.3	4.5	7.6
	−.1	5.1	**5.0**	5.1	5.3	5.3	4.9	6.4	4.4	6.6	**5.0**	5.1	5.6	5.7	5.4	4.9	7	4.5	7.9
	.1	5.3	**5.0**	5.1	5.1	**5.0**	5.1	6.6	4.4	6.4	5.4	**5.0**	5.6	5.7	5.2	5.1	6.9	4.4	7.6
	.5	5.3	5.4	**5.0**	**5.0**	4.4	5.1	6.7	4.3	5.9	5.3	5.4	5.3	5.3	4.6	**5.0**	6.6	4.5	7
	.9	5.2	**5.0**	5.3	4.1	4.6	4.8	5.9	4.3	5.1	**5.2**	5.3	5.6	6.5	3.3	5.4	7.6	4.3	7.6

Note. For each setting, the values closest to the prescribed level are printed in **bold** and values exceeding the upper limit (6.8%) of the 99% binomial interval are in red color.

Moreover, the bootstrapped test that is based on the maximum likelihood estimator T_L^* tends to behave similar to our three suggested bootstrap procedures in controlling type-I error rate. Only in the case of large positive correlation factors $\rho = .9$, it results in very conservative decisions.

In contrast, the other tests (T_N, T_F, T_P) do not control type-I error level constantly under heteroskedasticity or even under homoskedasticity in all of the considered sample sizes. It can also be seen from Tables 1, S.1, and S.2 that the NCT T_N, controls type-I error quite accurately in the case of larger numbers of complete pairs ($n_c = 30$), but it shows liberal behavior for smaller numbers of complete pairs ($n_c = 10$). This test turns very liberal in the case of heteroskedasticity. Furthermore, the FPM test T_F tends to result in a quite accurate type-I error control in the case of smaller numbers of complete pairs. For larger numbers of complete pairs, it leads to a conservative decision. For these scenarios, this behavior does not depend on the homoskedasticity assumption. Moreover, the NPC T_P shows a quite liberal behavior in most of the considered settings. Regarding the paired t-test based on the complete observations T_t, an inflation of the type-I error rate could be realized for certain distributions, when the missing rate was large and the number of complete pairs was small, see for example, the scenario $(n_c, n_u) = (10, 10)$. The effect vanishes for a larger number of complete pairs. This is in line with the theoretical results of the paired t-test with i.i.d. observations. The results also indicate that the paired bootstrapped t-test on the full data F controls type-I error through almost all settings.

It was also interesting to discover the type-I error rate control of the tests under *similar attributes to the breast cancer gene study data* which reflects data sets with a few pairs and large amount of unpaired portions. Simulation results for the type-I error rate of the studied procedures for $(n_c = 16, n_u = 74)$ sample sizes are presented in Tables S.22 and S.23. The correlation ρ in Table S.23 is estimated based on the data. It can be easily seen from Tables S.22 and S.23 that the bootstrap tests are

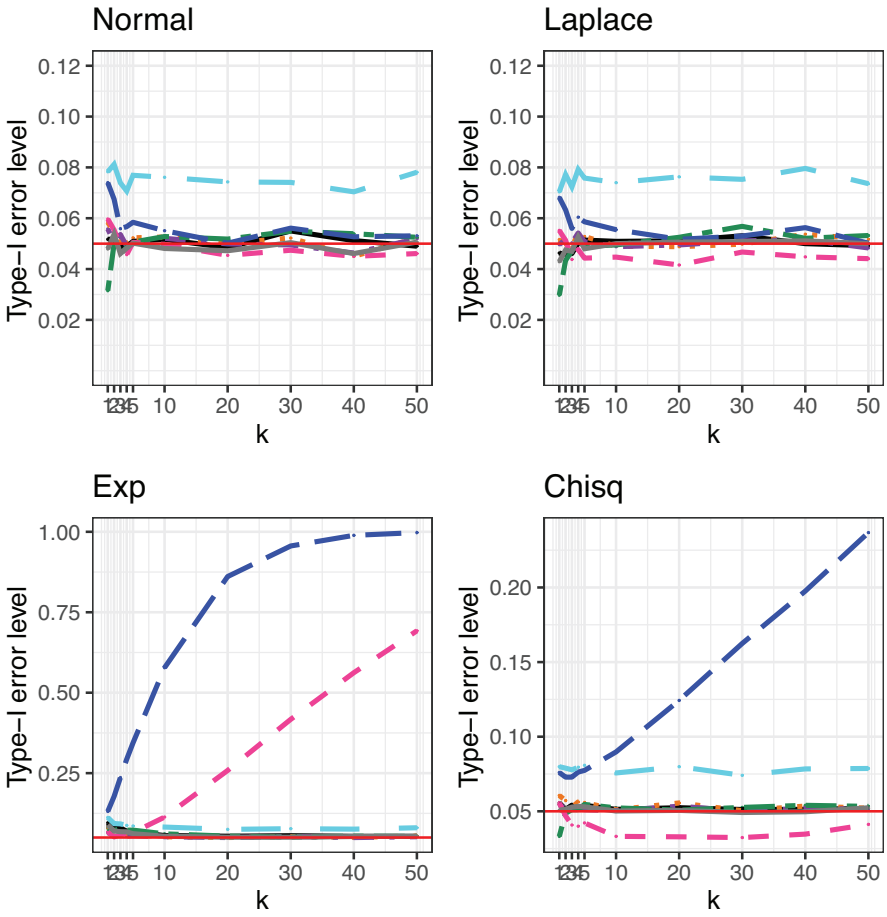

FIGURE 2 Type-I error simulation results ($\alpha = .05$) of the tests T_W^* (——), T_A^* (. . .), T_M^* (– · –), T_L^* (– – –), T_N (— —), T_F (– – –), T_t (—— ——), and T_P (– – –) for different distributions under correlation factor ($\rho = .9$) and heteroskedastic covariance matrix Σ_2 for varying k values multiplied by $(n_c, n_u) = (10, 30)$ under the MCAR framework

robust under large amounts of missing observations and control type-I error rate accurately, especially the bootstrapped tests T_W^*, and T_A^*. Except in the case of exponential distribution. The alternative approach T_N has acceptable control under homoskedasticity. But, under the exponential distribution, it turned very liberal especially under heteroskedasticity, while the Fisher's pooling method tends to result in quite acceptable control in most cases.

Simulation results of the type-I error level of the studied procedures under the MAR framework for different sample sizes and covariance structures are summarized in Tables S.3– S.5. There, it can be seen that for moderate to large sample sizes ($n \in \{20, 30\}$), the bootstrapped ATS T_A^*, the bootstrapped WTS T_W^*, the bootstrapped MATS T_M^*, the bootstrapped Little T_L^*, and the NCT T_N exhibit a fairly good type-I error rate control for almost all considered scenarios under homoskedasticity as well as heteroskedasticity. Only in the case of the skewed exponential distribution, the control of T_W^*, T_M^*, and T_N is not adequate and liberal behavior is observed, which is marked with red through all tables. In contrast, the bootstrapped MATS T_M^* tends to be sensitive to the dependency structure in the data. In particular, T_M^* exhibits a liberal behavior for negative correlations. For small sample sizes ($n = 10$), the T_N test tends to be liberal in all considered situations. In contrast, the bootstrapped tests T_W^*, T_M^*, and T_L^* exhibit good type-I error rate control for most settings except for the Laplace distribution. The bootstrapped ATS T_A^* tends to be very conservative especially under heteroskedasticity. However, the FPM T_F exhibits a conservative behavior under most considered situations.

Further Investigations on Type-I Error. In addition to the small and moderate sample size settings, we were also interested in studying type-I error rate control when *sample sizes increase*, while missing rates remain nearly unchanged. For moderate to large sample sizes, we considered the choices $(n_c, n_u) = k \cdot (10, 30)$ and $(n_c, n_u) = k \cdot (1, 1) + (10, 10)$, where k ranges from 1 to 50 (balanced case) and 0 to 500 (unbalanced case), respectively. Figures 2 and S.1 summarize the type-I error rate ($\alpha = .05$) for these settings. The results indicate that the NCT by Qi et al. (2019) T_N controls type-I error rate quite accurate under symmetric distributions, however, it fails to control type-I error rate unde skewed distributions.

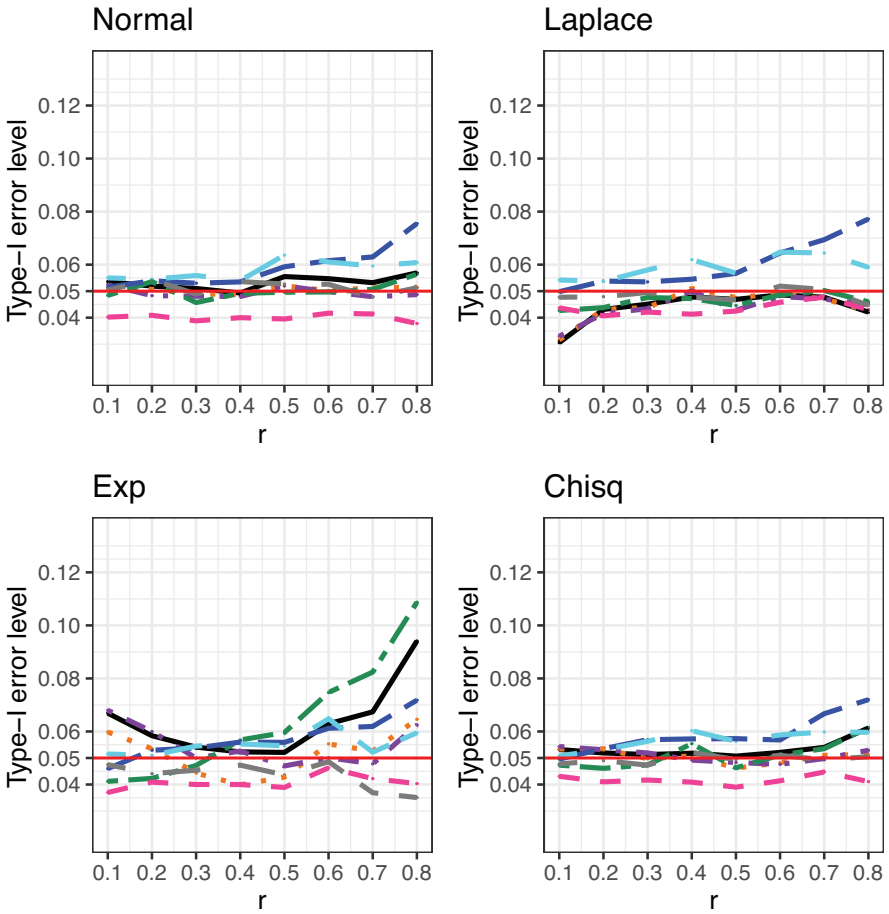

FIGURE 3 Type-I error simulation results ($\alpha = .05$) of the tests T_W^* (——), T_A^* (. . .), T_M^* (–·–), T_L^* (– – –), T_N (— —), T_F (– – –), T_t (—— ——), and T_P (– – –·) for different distributions under correlation factor ($\rho = .5$) with sample size ($n = 30$) and homoskedastic covariance matrix Σ_1 for varying missing rates $r \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$ under the MCAR framework

In fact, it gets even more liberal with increasing sample sizes. In contrast, the FPM T_F by Qi et al. (2019) tends to be conservative when missing rates among subjects of 50% are present, even under large numbers of complete observations such as $n_c = 510$ (Figure S.1). For larger missing rates (75%), it shows surprisingly quite accurate type-I error control (see Figure 2). Only in case of the exponential distribution, a very liberal behavior is observed that is acting analogous to a power function with increment of sample sizes (Figure 1). Here, the suggested bootstrap approaches T_A^*, T_W^* T_M^*, and T_L^* are the only methods that control type-I error rate accurately among all considered settings. The t-test T_t based on the complete cases controls type-I error as well, but had challenges with small complete cases $n_c \leq 10$. The NPC-test T_P, however, revealed a constant inflation of the type-I error rate for all missing rate scenarios. The degree of inflation remained the same even for increasing missing rates. Therefore, T_P seems not to be an adequate choice, even for smaller missing rates.

In order to cover the *effect of increasing missing rates*, we studied type-I error control for sample sizes of the form $(n_c, n_u) = ((1 - r) \cdot 30, r \cdot 30)$ with $r \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$ covering missing rates (among subjects) from 10% to 80% under moderate positive correlation factor ($\rho = .5$). Figures 3 and S.2 summarize type-I error rate control for these settings under a homoskedastic and a heteroskedastic covariance structure, respectively. The results indicate that under homoskedasticity, the alternative approach T_N tends to be slightly liberal. It moves closer to the 0.05 threshold for missing rates below 60%. In contrast, under heteroskedasticity, T_N tends to be more sensitive to the missing rates. In particular, it exhibits a conservative or liberal behavior for lower and larger missing rates, respectively. However, under this moderate sample size ($n = 30$) and correlation factor ($\rho = .5$), the FPM T_F tends to be conservative under all considered settings and its behavior is independent of the missing rate or even homoskedacticity assumption. In contrast, the suggested bootstrap approaches tend to control type-I error rate more accurate over the range of missing rates r for most settings. Only in case

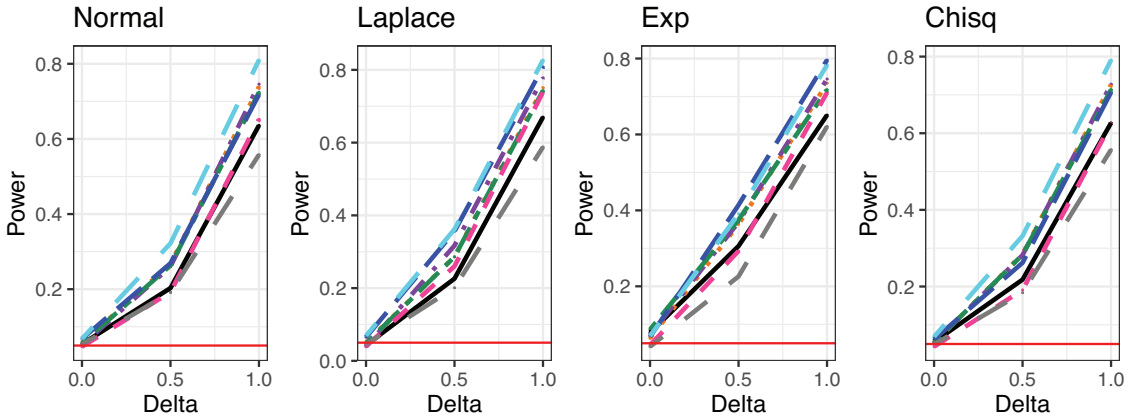

FIGURE 4 Power simulation results ($\alpha = .05$) of the tests T_W^* (——), T_A^* (. . .), T_M^* (– · –), T_L^* (– – –), T_N (— —), T_F (– – –), T_t (—— ——), and T_P (– – –.) for different distributions under correlation factor ($\rho = .1$) with sample size $(n_c, n_u) = (10, 30)$ and homoskedastic covariance matrix Σ_1 under the MCAR framework

of the skewed exponential distribution and missing rates greater than 50%, the control is not adequate. However, in this case all the other chosen procedures also failed to control the type-I error rate.

Power. In addition to the type-I error rate control, we studied the power of the nine tests for all considered settings. Figure 4 summarizes the power simulation results for a negative balanced sample size (10,30) under the MCAR framework. The power simulation results for the other scenarios are included in the supplement. The power analysis results of the considered methods under MCAR and MAR frameworks involving homoskedastic as well as heteroskedastic settings are summarized in Tables S6–S11 in supplement for the MCAR mechanism and Tables S12–S17 in supplement for the MAR mechanism. The entries that belong to very liberal tests have been colored in red in the power tables. It can be readily seen that the four bootstrapped tests T_W^*, T_A^*, T_M^*, and T_L^* and the NCT T_N have almost similar large power behavior under homoskedastic as well as heteroskedastic settings. Only in the heteroskedastic cases with skewed exponential distribution, the NCT T_N shows larger power than the others, which is due to its rather liberal behavior. One should also notice that the power behavior of each test varies based on the dependency structure of the data except for the bootstrapped ATS T_A^*. As expected, the paired t-test based on complete observations T_t revealed for small complete observations low power results compared to the other approaches. The NPC-test T_P also shows larger power results, but the effect can be leaded back to its liberal type-I error behavior.

7 | ILLUSTRATIVE DATA ANALYSES

In this section, we consider three real life problems coming from different sectors and sources. We start with a genome study on breast cancer.

7.1 | Breast cancer study: gene expression data

The TCGA project is a pilot project which was launched in 2005 with a financial support from the National Institutes of Health. It aims to understand the genetic basis of several types of human cancers through the application of high-throughput genome analysis techniques. TCGA collects molecular information such as miRNA/mRNA expressions, protein expressions, and weight of the sample as well as clinical data about the patients.

A breast cancer study has been performed by TCGA to improve the ability of diagnosing, treating, and preventing breast cancer through investigating the genetic basis of carcinoma. Their study consists of 1093 breast cancer patients with Clinical and RNA sequencing records. Among them, there were 112 subjects that provided both, normal, and tumor tissues. Here, we were interested in a subset of this datum that contains patients with pathologic stage I. This subset contains a total of $n_c = 16$ complete pairs and an unpaired sample for the patients who developed only tumor tissues of size $n_u = 74$. The data can be downloaded from Firehouse (www.gdac.broadinstitute.org).

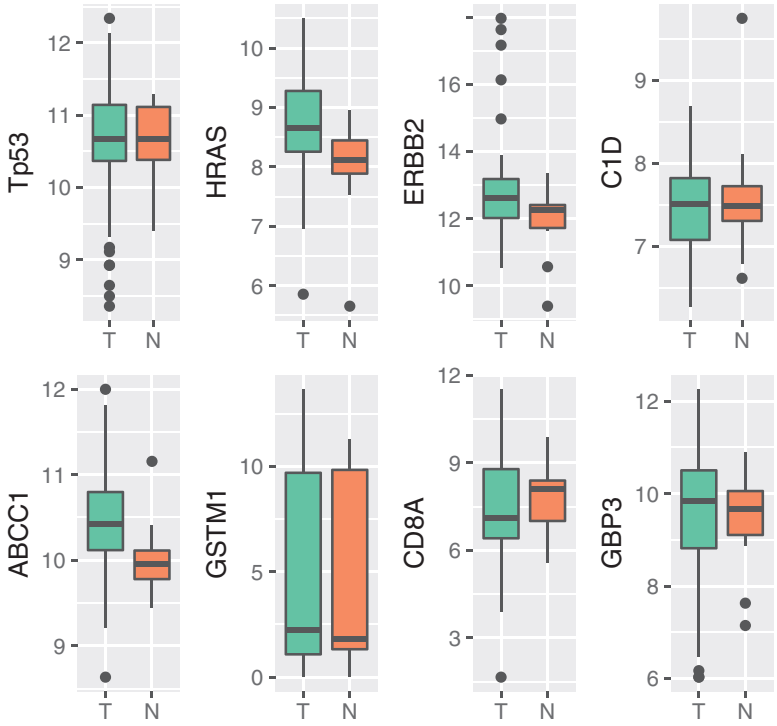

FIGURE 5 Profile of the gene expression levels of the tumor and normal breast tissues

TABLE 2 Unadjusted two-sided p-values of the breast cancer study

	Parametric bootstrap				Alternatives			
Gene	T^*_W	T^*_A	T^*_M	T^*_L	T_t	T_N	T_F	T_P
TP53	0.928	0.852	0.903	0.877	0.689	0.954	0.949	0.901
ABCC1	0.002	0.003	0.002	0.002	0.365	0.003	0.004	0
HRAS	0.007	0.002	0.003	0.002	0.022	0.001	0.004	0
GSTM1	0.821	0.85	0.849	0.515	0.605	0.629	0.967	0.827
ERBB2	0.043	0.024	0.011	0.014	0.136	0.071	0.069	0.007
CD8A	0.463	0.51	0.484	0.434	0.885	0.555	0.468	0.53
C1D	0.772	0.553	0.622	0.555	0.553	0.587	0.792	0.608
GBP3	0.196	0.301	0.214	0.084	0.083	0.103	0.357	0.262

Based on previous studies, six genes have been found to be significantly associated with breast cancer: **TP53, ABCC1, HRAS, GSTM1, ERBB2**, and **CD8A** (Harari & Yarden, 2000; De Jong et al., 2002; Munoz et al., 2007; Finak et al., 2008). Another two genes: **C1D** and **GBP3** were under investigation although they did not show any significant relation toward breast cancer patients (Qi et al., 2019). In this paper, we aim to test the hypothesis whether mean genetic expressions of the eight genes are significantly different between normal and tumor tissues for patients with early stage I breast cancer. Boxplots representing the characteristics of the eight genes are shown in Figure 5.

We applied all bootstrap testing methods T^*_W, T^*_A, T^*_M, and T^*_L as well as the alternative approaches T_t, T_N, T_F, and T_P to detect the null hypothesis of equal means between normal and tumor tissues ($H_0 : \mu_1 = \mu_2$) against the two-sided alternative ($H_1 : \mu_1 \neq \mu_2$). The results are summarized in Table 2.

It can be seen from Table 2 that the bootstrapped approaches T^*_W, T^*_A, T^*_M, and T^*_L and the NPC T_P identified three of eight genes having significantly different genetic expressions in normal and tumor tissues; genes **ABCC1, HRAS**, and **ERBB2**. However, the NCT T_N, and the FPM T_F led to different results for the **ERBB2** gene. Regarding the paired t-test based on the complete observations T_t, different results obtained for the **ABCC1** and **ERBB2** genes.

TABLE 3 Two-sided p-values of the considered studies

		Parametric bootstrap				Alternatives			
Study	F	T^*_{W}	T^*_{A}	T^*_{M}	T^*_{L}	T_t	T_{N}	T_{F}	T_{P}
Anorexia	0.002	0.026	0.043	0.029	0.03	0.004	0.136	0.008	0.022
GrapeFruit	0.002	0.039	0.014	0.029	0.068	0.141	0.031	0.068	0.022

7.2 | Two more examples

To illustrate potential differences between all methods we consider two additional examples called "Anorexia" and "GrapeFruit." Each of them consists of complete data sets and missing values were introduced on them by the MCAR mechanism with a missing rate of $r = 30\%$. They can be briefly described as follows:

Anorexia. This data set consists of weights in pounds for 17 young girls who were receiving a treatment for anorexia over a fixed period of time. The main problem is to compare the girls' weights before and after the treatment. This datum was originally published by Hand et al. (1993), and were analyzed in Pruzek & Helmreich (2009). It is also included in the R package PairedData (Champely & Champely, 2018).

GrapeFruit. It consists of a paired samples data that are taken from Preece (1982). The study aimed to detect differences between "shaded" and "exposed" grapefruits. To make the differences as precise as possible, they dealt with both halves of a single fruit under similar conditions. This data set consists of the percentages of solids in the shaded and exposed halves of 25 grapefruits. This datum is also contained in the R package PairedData (Champely & Champely, 2018).

We applied the F-test that considers the full data before missingness, all bootstrapped approaches T^*_W, T^*_A, T^*_M, and T^*_L as well as the alternative approaches T_t, T_N, T_F, and T_P to detect the null hypothesis of equal means $H_0 : \{\mu_1 = \mu_2\}$ against the two-sided alternative $H_1 : \{\mu_1 \neq \mu_2\}$. The results are summarized in Table 3. It can be seen from Table 3 that the full data test F, bootstrapped approaches T^*_W, T^*_A, and T^*_M, and the Pesarin test T_P identified significant differences in both data sets. However, the alternative naive approach based on the complete observations T_t and the FPM T_F failed in detecting significant difference in the GrapeFruit data set. In addition, the NCT T_N could not identify any significant difference for the Anorexia data set.

8 | DISCUSSION AND OUTLOOK

The problem of matched pairs with missing values occurs frequently in practice. Most available procedures in the literature are not applicable when missing values occur in a single arm. Exceptions are given by the recent NCT and FPM approaches of Qi et al. (2019). For the NCT approach, Qi et al. (2019) utilize a combination of the sign and Wilcoxon Mann–Whitney rank sum test. And, the FPM approach, is based on a weighted combination of the p-values of the Wilcoxon signed rank test and the Wilcoxon Mann–Whitney rank sum test. For homoskedastic settings with symmetric distributions, the NCT and FPM approaches can be recommended. If, however, the underlying assumptions are not true (e.g., in skewed heteroskedastic setups), the NCT and FPM may result in highly inflated type-I errors or considerable power loss. In addition to the NCT and FPM approaches, we also studied a single-arm missingness modification of a nonparametric testing procedure given in Pesarin & Salmaso (2010). It is based on the usage of the permutation paired t-test and the permutation Welch test on partial combination of the whole data $\mathcal{D}_n$ with missingness. However, the proposed combination strategy did not reveal favorable results leading to a constant inflation of the type-I error. We also calculated the paired t-test based on complete observations only.

To overcome all these issues, we have provided resampling procedures that are not based on any parametric assumptions and use all observed information within the matched pairs design. They were shown to be asymptotically correct and robust under heteroskedasticity and skewed distributions. The tests were based on restructuring all observed information in a test statistic of quadratic form that can be either a WTS, an ATS, or a MATS. Since WTS is well known (from other situations like in Vallejo et al., 2010; Konietschke et al., 2015; or Smaga, 2017) for being liberal, while ATS and MATS tend to be rather conservative or liberal for small to moderate sample sizes, we improved their small sample behavior by an asymptotic model-based bootstrap approach. The procedure's asymptotic validity was also proven and can be found in the supplement. In addition, we improved the behavior of the Little's test (cf. Little, 1976) that is based upon a modified maximum likelihood estimator by introducing an asymptotic model-based bootstrap version of the test.

In an extensive simulation study, the type-I error rate control of the tests have been examined for symmetric and skewed distributions with homoskedastic and heteroskedastic covariance settings under different missing mechanisms. There, it was seen that the parametric bootstrap versions of WTS, ATS, MATS, and Little improve their small sample behavior. In particular, our bootstrap tests have been shown to perform very well in most of the cases, even with larger amount of missingness, heteroskedastic covariance or skewed data. Only the type-I error control for the exponential distribution, particularly under heteroskedasticity, MCAR and small paired sample sizes with rather large unpaired portions ($n_c = 10, n_u = 30$), is not maintained. In this setting, however, all other considered methods such as the ones given in Qi et al. (2019) and inspired by Pesarin (2001) and Pesarin & Salmaso (2010) also failed to control the type-I error rate.

Furthermore, our simulation study reveals that the bootstrap procedures' type-I error control is not much affected by less stringent missing data mechanism such as the MAR. However, their power behavior is affected. A possible justification of the latter effect might originate from the additional dependence structure within the occurrence of missing values compared to the MCAR case. It seems that the testing procedure is more challenged to detect deviations from the null.

In order to simplify the application of our approaches, the three proposed bootstrap statistical methods have been implemented within the PBT function in the freely available R-package **MissPair**. It is available on GitHub (https://github.com/lubnaamro/MissPair) and will be available on the CRAN repository.

Future research will be concerned with extending our procedures to multivariate settings (MANOVA). An investigation of the behavior of our methods together with *logit* or *probit* transformations may also be part of future work.

ACKNOWLEDGMENTS

Burim Ramosaj and Markus Pauly acknowledge the support of the German Research Foundation (DFG). Lubna Amro's work was also supported by the German Academic Exchange Service (DAAD) under the project: Research Grants - Doctoral Programmes in Germany, 2015/16 (No. 57129429).

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that supports the findings of this study are available in the supplementary material 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 could fully be reproduced.

ORCID

Lubna Amro https://orcid.org/0000-0001-8550-7406
Markus Pauly https://orcid.org/0000-0002-0976-7190
Burim Ramosaj https://orcid.org/0000-0002-1885-5168

REFERENCES

Amro, L., Konietschke, F., & Pauly, M. (2019). Multiplication-combination tests for incomplete paired data. *Statistics in Medicine*, *38*, 3243–3255.

Amro, L., & Pauly, M. (2017). Permuting incomplete paired data: A novel exact and asymptotic correct randomization test. *Journal of Statistical Computation and Simulation*, *87*, 1148–1159.

Bhoj, D. S. (1978). Testing equality of means of correlated variates with missing observations on both responses. *Biometrika*, *65*, 225–228.

Brown, M. B. (1975). 400: A method for combining non-independent, one-sided tests of significance. *Biometrics*, *31*, 987–992.

Brunner, E. (2001). Asymptotic and approximate analysis of repeated measures designs under heteroscedasticity. *Mathematical Statistics with Applications in Biometry*, *Festschrift in Honour of Prof. Dr. Siegfried Schach*, 313–326.

Champely, S., & Champely, M. S. (2018). R-Package "PairedData" Version 1.1.1.

Chung, E., & Romano, J. P. (2013). Exact and asymptotically robust permutation tests. *Annals of Statistics*, *41*, 484–507.

De Jong, M., Nolte, I., Te Meerman, G., Van der Graaf, W., Oosterwijk, J., Kleibeuker, J., Schaapveld, M., & De Vries, E. (2002). Genes other than *BRCA1* and *BRCA2* involved in breast cancer susceptibility. *Journal of Medical Genetics*, *39*, 225–242.

Ekbohm, G. (1976). On comparing means in the paired case with incomplete data on both responses. *Biometrika*, *63*, 299–304.

Feng, Q., Hawes, S. E., Stern, J. E., Wiens, L., Lu, H., Dong, Z. M., Jordan, C. D., Kiviat, N. B., & Vesselle, H. (2008). DNA methylation in tumor and matched normal tissues from non-small cell lung cancer patients. *Cancer Epidemiology and Prevention Biomarkers*, *17*, 645–654.

Finak, G., Bertos, N., Pepin, F., Sadekova, S., Souleimanova, M., Zhao, H., Chen, H., Omeroglu, G., Meterissian, S., Omeroglu, A., Hallett, M., & Park, M.. (2008). Stromal gene expression predicts clinical outcome in breast cancer. *Nature Medicine*, *14*, 518–527.

Friedrich, S., & Pauly, M. (2018). Mats: Inference for potentially singular and heteroscedastic MANOVA. *Journal of Multivariate Analysis*, *165*, 166–179.

Fuchs, N., Pölz, W., & Bathke, A. C. (2017). Confidence intervals for population means of partially paired observations. *Statistical Papers*, *58*, 35–51.

Hand, D. J., Daly, F., McConway, K., Lunn, D., & Ostrowski, E. (1993). *A handbook of small data sets*. CRC Press.

Harari, D., & Yarden, Y. (2000). Molecular mechanisms underlying ErbB2/HER2 action in breast cancer. *Oncogene*, *19*, 6102–6114.

Harrar, S. W., Feyasa, M. B., & Wencheko, E. (2020). Nonparametric procedures for partially paired data in two groups. *Computational Statistics & Data Analysis*, *144*, 106903.

Hartung, J. (1999). A note on combining dependent tests of significance. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *41*, 849–855.

Hou, C.-D. (2005). A simple approximation for the distribution of the weighted combination of non-independent or independent probabilities. *Statistics & Probability Letters*, *73*, 179–187.

Janssen, A. (1997). Studentized permutation tests for non-i.i.d. hypotheses and the generalized Behrens-Fisher problem. *Statistics & Probability Letters*, *36*, 9–21.

Janssen, A. (1999). Nonparametric symmetry tests for statistical functionals. *Mathematical Methods of Statistics*, *8*, 320–343.

Kim, B. S., Kim, I., Lee, S., Kim, S., Rha, S. Y., & Chung, H. C. (2004). Statistical methods of translating microarray data into clinically relevant diagnostic information in colorectal cancer. *Bioinformatics*, *21*, 517–528.

Konietschke, F., Bathke, A. C., Harrar, S. W., & Pauly, M. (2015). Parametric and nonparametric bootstrap methods for general MANOVA. *Journal of Multivariate Analysis*, *140*, 291–301.

Konietschke, F., & Pauly, M. (2014). Bootstrapping and permuting paired t-test type statistics. *Statistics and Computing*, *24*, 283–296.

Kost, J. T., & McDermott, M. P. (2002). Combining dependent p-values. *Statistics & Probability Letters*, *60*, 183–190.

Krishnamoorthy, K., & Lu, F. (2010). A parametric bootstrap solution to the MANOVA under heteroscedasticity. *Journal of Statistical Computation and Simulation*, *80*, 873–887.

Kuan, P. F., & Huang, B. (2013). A simple and robust method for partially matched samples using the p-values pooling approach. *Statistics in Medicine*, *32*, 3247–3259.

Kuriakose, M. A., Chen, W. T., He, Z. M., Sikora, A. G., Zhang, P., Zhang, Z. Y., Qiu, W. L., Hsu, D. F., McMunn-Coffran, C., Brown, S. M., Elango, E. M., Delacure, M. D., & Chen, F. A., et al,. (2004). Selection and validation of differentially expressed genes in head and neck cancer. *Cellular and Molecular Life Sciences: CMLS*, *61*, 1372–1383.

Lapointe, J., Li, C., Higgins, J. P., Van De Rijn, M., Bair, E., Montgomery, K., Ferrari, M., Egevad, L., Rayford, W., Bergerheim, U., Ekman, P., DeMarzo, A., Tibshirani, R., Botstein, D., Brown, P., Brooks, J., & Pollack, J.. (2004). Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proceedings of the National Academy of Sciences*, *101*, 811–816.

Lin, P.-E. (1973). Procedures for testing the difference of means with incomplete data. *Journal of the American Statistical Association*, *68*, 699–703.

Lin, P.-E., & Stivers, L. E. (1974). On difference of means with incomplete data. *Biometrika*, *61*, 325–334.

Little, R. J. (1976). Inference about means from incomplete multivariate data. *Biometrika*, *63*, 593–604.

Looney, S. W., & Jones, P. W. (2003). A method for comparing two normal means using combined samples of correlated and uncorrelated data. *Statistics in Medicine*, *22*, 1601–1610.

Martínez-Camblor, P., Corral, N., & María de la Hera, J. (2013). Hypothesis test for paired samples in the presence of missing data. *Journal of Applied Statistics*, *40*, 76–87.

Mehta, J., & Gurland, J. (1969). Testing equality of means in the presence of correlation. *Biometrika*, *56*, 119–126.

Mehta, J., & Gurland, J. (1973). A test for equality of means in the presence of correlation and missing values. *Biometrika*, *60*, 211–213.

Morrison, D. F. (1973). A test for equality of means of correlated variates with missing data on one response. *Biometrika*, *60*, 101–105.

Munoz, M., Henderson, M., Haber, M., & Norris, M. (2007). Role of the MRP1/ABCC1 multidrug transporter protein in cancer. *IUBMB Life*, *59*, 752–757.

Pauly, M., Brunner, E., & Konietschke, F. (2015). Asymptotic permutation tests in general factorial designs. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, *77*, 461–473.

Pesarin, F. (2001). *Multivariate permutation tests: With applications in Biostatistics*. Wiley Chichester.

Pesarin, F., & Salmaso, L. (2010). *Permutation tests for complex data: Theory, applications and software*. Wiley.

Preece, D. (1982). T is for trouble (and textbooks): A critique of some examples of the paired-samples t-test. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *31*, 169–195.

Pruzek, R. M., & Helmreich, J. E. (2009). Enhancing dependent sample analyses with graphics. *Journal of Statistics Education*, *17*.

Qi, Q., Yan, L., & Tian, L. (2019). Testing equality of means in partially paired data with incompleteness in single response. *Statistical Methods in Medical Research*, *28*, 1508–1522.

Ramosaj, B., Amro, L., & Pauly, M. (2020). A cautionary tale on using imputation methods for inference in matched-pairs design. *Bioinformatics*, *36*, 3099–3106.

Ramosaj, B., & Pauly, M. (2019). Predicting missing values: A comparative study on non-parametric approaches for imputation. *Computational Statistics*, *34*, 1741–1764.

Rempala, G. A., & Looney, S. W. (2006). Asymptotic properties of a two sample randomized test for partially dependent data. *Journal of Statistical Planning and Inference*, *136*, 68–89.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys* (Vol. 81). Wiley.

Samawi, H. M., Helu, A., & Vogel, R. (2011). A nonparametric test of symmetry based on the overlapping coefficient. *Journal of Applied Statistics*, *38*, 885–898.

Samawi, H. M., & Vogel, R. (2014). Notes on two sample tests for partially correlated (paired) data. *Journal of Applied Statistics*, *41*, 109–117.

Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research*, *8*, 3–15.

Smaga, Ł. (2017). Bootstrap methods for multivariate hypothesis testing. *Communications in Statistics-Simulation and Computation*, *46*, 7654–7667.

Stekhoven, D. J., & Bühlmann, P. (2011). Missforest—Non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*, 112–118.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ*, *338*, b2393.

Uddin, N., & Hasan, M. (2017). Testing equality of two normal means using combined samples of paired and unpaired data. *Communications in Statistics - Simulation and Computation*, *46*, 2430–2446.

Vallejo, G., Fernández, M., & Livacic-Rojas, P. E. (2010). Analysis of unbalanced factorial designs with heteroscedastic data. *Journal of Statistical Computation and Simulation*, *80*, 75–88.

Van Buuren, S. (2018). *Flexible imputation of missing data*. Chapman and Hall/CRC.

Waljee, A. K., Mukherjee, A., Singal, A. G., Zhang, Y., Warren, J., Balis, U., Marrero, J., Zhu, J., & Higgins, P. D. (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open*, *3*, e002847.

Xu, J., & Harrar, S. W. (2012). Accurate mean comparisons for paired samples with missing data: An application to a smoking-cessation trial. *Biometrical Journal*, *54*, 281–295.

Xu, L.-W., Yang, F.-Q., Abula, A., & Qin, S. (2013). A parametric bootstrap approach for two-way ANOVA in presence of possible interactions with unequal variances. *Journal of Multivariate Analysis*, *115*, 172–180.

Yu, D., Lim, J., Liang, F., Kim, K., Kim, B. S., & Jang, W. (2012). Permutation test for incomplete paired data with application to CDNA microarray data. *Computational Statistics & Data Analysis*, *56*, 510–521.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Amro L, Pauly M, Ramosaj B. Asymptotic-based bootstrap approach for matched pairs with missingness in a single arm. *Biometrical Journal*. 2021;*63:*1389–1405. https://doi.org/10.1002/bimj.202000051