Research Article

Statistics in Medicine

Received: 12 January 2016, Accepted: 31 January 2017 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.7263

Analysis of small sample size studies using nonparametric bootstrap test with pooled resampling method

Alok Kumar Dwivedi,[a,b*†] Indika Mallawaarachchi[b] and Luis A. Alvarado[b]

Experimental studies in biomedical research frequently pose analytical problems related to small sample size. In such studies, there are conflicting findings regarding the choice of parametric and nonparametric analysis, especially with non-normal data. In such instances, some methodologists questioned the validity of parametric tests and suggested nonparametric tests. In contrast, other methodologists found nonparametric tests to be too conservative and less powerful and thus preferred using parametric tests. Some researchers have recommended using a bootstrap test; however, this method also has small sample size limitation. We used a pooled method in nonparametric bootstrap test that may overcome the problem related with small samples in hypothesis testing. The present study compared nonparametric bootstrap test with pooled resampling method corresponding to parametric, nonparametric, and permutation tests through extensive simulations under various conditions and using real data examples. The nonparametric pooled bootstrap *t*-test provided equal or greater power for comparing two means as compared with unpaired *t*-test, Welch *t*-test, Wilcoxon rank sum test, and permutation test while maintaining type I error probability for any conditions except for Cauchy and extreme variable lognormal distributions. In such cases, we suggest using an exact Wilcoxon rank sum test. Nonparametric bootstrap paired *t*-test also provided better performance than other alternatives. Nonparametric bootstrap test provided benefit over exact Kruskal–Wallis test. We suggest using nonparametric bootstrap test with pooled resampling method for comparing paired or unpaired means and for validating the one way analysis of variance test results for non-normal data in small sample size studies. Copyright © 2017 John Wiley & Sons, Ltd.

Keywords: bootstrap test; nonparametric test; parametric test; resampling method; small sample size; experimental studies

Introduction

Common designs in biomedical research include experimental designs in laboratory studies and pilot randomized controlled designs in clinical studies. Data from these studies are often analyzed using simple univariate statistical tests. In such studies, some methodologists have suggested using parametric tests, whereas others have preferred the use of nonparametric tests [1,2]. Generally, these studies are based on small sample sizes, thus conventional statistical guidelines would recommend using nonparametric approaches for analyzing data from these studies [3]. There are two ways of obtaining *p*-values in nonparametric tests. One way calculates exact probability of obtaining observed or more extreme results under the null hypothesis, which is suitable for small sample size studies, and referred to as exact nonparametric test. The other way calculates *p*-value based on asymptotic property, which is suitable for large sample size studies, and referred to as asymptotic nonparametric test. Asymptotic and exact procedures for computing *p*-values for most of the nonparametric tests are available. These tests are useful when the assumptions of parametric tests are under suspicion [4,5]. However, standard exact or asymptotic nonparametric methods do not perform well in many conditions with small sample size

[a] *Division of Biostatistics and Epidemiology, Department of Biomedical Sciences, Paul L. Foster School of Medicine, Texas Tech University Health Sciences Center, El Paso, Texas, U.S.A.*

[b] *Biostatistics and Epidemiology Consulting Lab, Office of Research Resources, Texas Tech University Health Sciences Center, El Paso, Texas, U.S.A.*

**Correspondence to: Division of Biostatistics and Epidemiology, Department of Biomedical Sciences, Paul L. Foster School of Medicine, Texas Tech University Health Sciences Center, El Paso, Texas, U.S.A.*

[†] *E-mail: alok.dwivedi@ttuhsc.edu*

studies. The other alternative approaches for small sample size studies are resampling methods such as nonparametric permutation and bootstrap tests. In the permutation method, the test statistic values are obtained in all possible permutation resamples (without replacement), whereas the test statistic values are obtained in bootstrap resamples (with replacement) in the bootstrap test under the null hypothesis. Permutation test may not perform well in small sample size studies as it provides only a few permutations. Barber and Thompson recommended using a bootstrap technique for either checking the robustness of parametric methods or for primary statistical analysis for comparing means in moderate-sized or large-sized studies with skewed data [6]. Another study assessed the properties of the bootstrap test and showed that bootstrap tests are reliable for a sample as small as eight [7]. The performance of bootstrap test as compared with common parametric and nonparametric including permutation tests has not been studied well in small sample size studies.

The key assumption of any parametric tests is normal distribution. Some parametric tests require a homogeneity of variances assumption. If such assumptions are not satisfied, then it may increase type I error (false positive: inappropriately reject the null hypothesis when there is no difference) or increase the type II error (not rejecting the null hypothesis when there is a difference), and subsequently reduce power (1-probability of type II error). Thus, the choice of statistical tests (parametric versus nonparametric) is made on the basis of relative to these errors [8]. There are conflicting findings reported regarding the use of parametric and nonparametric tests. Some studies demonstrated greater statistical power for nonparametric tests compared with parametric tests in small sample size studies and with non-normal distributions [9–11]. On the contrary, other studies suggested that nonparametric tests have less or even no power, in small sample size studies; and thus, a parametric test should be used [2,12–14]. Segal [1] argued that it is inappropriate to assess a normality assumption in small sample size studies; therefore, nonparametric tests are the only choice and should be used in analysis of small sample size studies. These conflicting findings arose due to two reasons: (1) central limit theorem, as per the central limit theorem, if the sample size is reasonably large, then parametric tests can be used even if the original distribution of data is non-normal. However, the threshold of sample size considered to be large or small in order to apply central limit theorem is unclear. It has been noticed that if the distribution closely resembles a normal distribution, then sampling distribution of mean would approximately follow the normal distribution for sample size of 5–10, while for other distributions, a sample size of at least 30 requires us to follow the normal distribution of the sampling distribution of mean [15,16]. In some extreme skewed distributions, a sample size of 100, 500, or 1000 or more was required to achieve normal distribution [15,17,18]; (2) sample size and statistical power for nonparametric tests, Siegel and Castellan provided threshold values above which asymptotic procedure of nonparametric tests is permitted [4]. Accordingly, at least 15 or more data points are needed to conduct common asymptotic nonparametric tests. Further, these authors suggested using exact testing if the sample size is extremely small [4,5]. Sokal and Rohlf [19] suggest the use of the exact test for sample sizes <50. However, exact nonparametric tests cannot be permitted for extremely small sample size studies. Use of exact nonparametric tests may also provide insignificant *p*-values even with large differences in extremely small sample size studies. Thus, researchers started using parametric tests in situations where either nonparametric tests provide low power or original sample distribution is roughly normal, irrespective of sample size of the study or without checking normality or equality of variance assumption.

We need an alternative hypothesis testing method that requires minimum or even no assumptions related with the distribution, provides a relatively large or equal power to parametric tests, and controls type I error reasonably. One such alternative could be a nonparametric bootstrap test. The two bootstrap test versions (nonparametric or parametric) can be used for hypothesis testing. The nonparametric bootstrap test involves no assumptions related with underlying population distribution, and thus, can be a competitive alternative for hypothesis testing for small or extremely small sample size studies. Very limited studies have examined the feasibility of nonparametric bootstrap tests compared with other alternatives in small sample size studies [20–25]. Of these, very few studies have provided extensive simulations under various conditions. Moreover, most of these studies have used different test statistic or different resampling procedures for determining sampling distribution under null hypothesis for conducting bootstrap test, which end up with conflicting findings. In summary, there is a confusion regarding the use of nonparametric bootstrap test, parametric or nonparametric methods including permutation tests for comparing means between groups (paired/unpaired) especially for small sample size studies. Thus, there is a critical need to establish recommendations for using appropriate statistical tests in small sample size studies. Broadly, there are two resampling strategies: one is unpooled resampling strategy in which bootstrap sample is drawn from each marginal sample separately and other is pooled/mixed resampling strategy in which bootstrap sample is drawn from pooled sample. For small sample size studies, unpooled resampling approach provides less resampling variability [21]. We used a pooled resampling method to compute p-value using a nonparametric bootstrap test for comparisons of independent or dependent means. In this study, we summarize the minimum sample size needed of which exact procedure of nonparametric tests is permitted and assess the power and false positivity of proposed nonparametric bootstrap tests compared with their alternatives, standard parametric and nonparametric including permutation tests.

Statistical tests

Parametric tests

Parametric tests are often used for hypothesis testing and allow direct comparison of means. Common parametric tests are unpaired t-test, Welch t-test, paired t-test, and one way analysis of variance (F test). Unpaired t-test and one way analysis of variance require assumption of normality and equal variances, while Welch t-test and paired t-test require only the assumption of normality. The violation of these assumptions may seriously affect the validity of parametric tests. Parametric tests have been shown to be robust under violation of these assumptions for large sample size studies. Unfortunately, it is very hard to judge if standard parametric tests will be valid for non-normal or heteroscedastic data, particularly in small sample size studies and large sample size studies with non-normal data.

Standard nonparametric tests

Standard nonparametric tests have been used when assumptions of parametric tests cannot be achieved or the sample size is small. The most common nonparametric tests are Wilcoxon rank sum test, Wilcoxon signed-rank test, and Kruskal–Wallis test. These nonparametric methods compare the distributions between groups. Nonparametric tests have generally shown to be inferior to parametric tests when assumptions related with parametric tests are met. Significant result of a nonparametric test does not differentiate whether the difference is between the location and shape of the distributions. Thus, it limits the use of nonparametric tests especially where the shape of distribution between groups is very different.

Permutation tests

Standard nonparametric tests ignore distributional information by converting data to ranks, which may be undesirable for researchers or may produce less powerful results compared with parametric tests. The nonparametric permutation and bootstrap test avoid such controversies by retaining distributional information through original sample [21,26]. The idea of permutation method was introduced by R. A. Fisher [27]. Different methods of using permutation test, their advantages and disadvantages have been discussed [24,28–32]. The basic idea of a permutation test is to determine the sampling distribution of a test statistic by permuting the observations under the null hypothesis. The permutation test requires assumption of exchangeability [29], independent identical distributed data [33], and identical distributions with the same shapes and spreads [34]. However, a recent article suggested that permutation test does not necessarily require the assumption of exchangeability [35]. Three methods are commonly used for computing the p-value under permutation test: (1) an exact permutation test in which complete enumeration of permutations of the data is obtained to compute the p-value. This method is computationally intensive and preserves the type I error being exact test. It generally reduces the power of the test; (2) resampling or Monte Carlo permutation test in which permutation distribution is approximated by permuting data over a subset of all possible permutations. This exact method is also very computer-intensive for large sample size studies; (3) asymptotic permutation test in which the p-value is computed by following asymptotic normality of the permutation distribution of the test statistic. The asymptotic version of permutation test often provides larger power relative to exact permutation tests [28,36]. There is a debate regarding the use of permutation methods over the conventional statistical tests because of conflicting findings reported in the literature [9,37,38].

Bootstrap tests

Like permutation test, bootstrap test is also a data-based resampling statistical method and serves as a competitive approach for hypothesis testing. The bootstrap method was developed by Efron and Tibshirani [32]. This resampling procedure can be applied in two ways: (1) parametric bootstrap testing in which it is assumed that the data comes from a known distribution that is represented by a real data sample; (2) nonparametric bootstrap test in which it is assumed that the real data sample represents the empirical distribution of the population. From the assumed distribution or empirical distribution of population under the null hypothesis, a large number of bootstrap samples are drawn for constructing a sampling distribution of a test statistic. This is called as bootstrap sample distribution. Bootstrap test *p*-value is obtained by locating the observed statistic of interest from the observed sample on the bootstrap distribution. Nonparametric bootstrap test has been recommended for comparing means in the moderate to large sample size studies [6]; however, the feasibility of this method has been suspected in small sample size studies [20]. The performance of a bootstrap test depends on the choice of a test statistic and resampling approach to construct bootstrap distribution, which is described in next section.

Methods

Selection of test statistic in bootstrap test

In hypothesis testing using bootstrap test, selection of reliable test statistic is very important. Different choices of test statistic in the bootstrap test may provide different results [6]. We selected a studentized *t*-statistic [32] for comparing two means and an F-statistic [39] for comparing more than two means. When comparing two independent means, the test statistic is defined as

$T_b = \theta_b/SE(\theta_b)$, where θ_b is the difference in means in bootstrap sample under the null hypothesis and SE is the standard error of the difference in means in bootstrap sample.

Similarly, for comparing two dependent means, we selected *t*-statistic, which is defined as

$T_b = \theta_b/SE(\theta_b)$, where θ_b is the mean of change outcome in the bootstrap sample under the null hypothesis and SE is the standard error of the change outcome in bootstrap sample.

For comparing more than two independent means, we selected *F*-statistic, which is defined as

$F_b = MSB_b/MSW_b$, where MSB is the mean square between groups and MSW is the mean square within groups in bootstrap sample.

We referred bootstrap tests as nonparametric bootstrap *t*-test for comparing two independent means, nonparametric bootstrap *F*-test for comparing more than two independent means, and nonparametric bootstrap paired *t*-test for comparing two dependent means.

Methods for resampling under the null hypothesis

The resampling strategy under the null hypothesis is a necessary step in computing bootstrap test *p*-value. For comparing two independent means, Efron and Tibshirani suggested an algorithm 16.2, page 224 [32], referred as unpooled error bootstrap test, described as:

(1) Let $x = x_1, x_2\ldots, x_m$ is the observed sample 1 of size m with mean $\bar{x}$ and variance σ_x^2 and $y = y_1, y_2\ldots, y_n$ is the observed sample 2 of size n with mean $\bar{y}$ and variance σ_y^2.
(2) Evaluate test statistic [t(.)] such as $t_{obs} = \frac{\bar{x}\text{-}\bar{y}}{\sqrt{\sigma_x^2/m+\sigma_y^2/n}}$.
(3) Create two transformed error datasets $x^* = x_1 - \bar{x} + \bar{z},\ \ x_2 - \bar{x} + \bar{z}, \ldots\ldots x_m - \bar{x} + \bar{z}$ and $y^* = y_1 - \bar{y} + \bar{z},\ \ y_2\text{-}\bar{y} + \bar{z}, \ldots\ldots y_n - \bar{y} + \bar{z}$, where $\bar{z}$ is the mean of the combined sample.
(4) Draw two bootstrap samples: one of size m observations with replacement ($x^{*\prime}$) and another of size n observations with replacement ($y^{*\prime}$). Compute mean and variance of each bootstrap sample as ($\bar{x}^{*\prime}, \sigma_x^{2*\prime}$) and ($\bar{y}^{*\prime}, \sigma_y^{2*\prime}$), respectively.
(5) Evaluate test statistic: $t^{*\prime} = \frac{\bar{x}^{*\prime} - \bar{y}^{*\prime}}{\sqrt{\sigma_x^{2*\prime}/m+\sigma_y^{2*\prime}/n}}$.
(6) Repeat steps 3 and 4 for B (e.g., 1000) times and obtain B values of the test statistic ($t^{*\prime}$).
(7) Approximate *p*-value=$\frac{\text{number of times } (t^{*\prime} \geq t_{obs})}{B}$.

Lunneborg proposed pooled error bootstrap and unpooled error bootstrap procedures [40]. The pooled error bootstrap procedure is computed as follows:

(1) Let $x = x_1, x_2\ldots, x_m$ is the observed sample 1 of size m with mean $\bar{x}$ and variance σ_x^2 and $y = y_1, y_2\ldots, y_n$ is the observed sample 2 of size n with mean $\bar{y}$ and variance σ_y^2.

(2) Evaluate test statistic as $t_{obs} = \frac{\overline{x}-\overline{y}}{\sqrt{\sigma_x^2/m+\sigma_y^2/n}}$.

(3) Create two transformed error datasets $x^* = x_1-\bar{x}, x_2-\bar{x} \ldots, x_m-\bar{x}$ and $y^* = y_1-\bar{y}, y_2-\bar{y} \ldots, y_n-\bar{y}$.

(4) Draw two bootstrap samples: one of size m observations with replacement (x**) and another of size n observations with replacement (y**) from the pooled error sample (referred as pooled error bootstrap test). Compute mean and variance of each bootstrap sample as $(\overline{x}^{**}, \sigma_x^{2**})$ and $(\overline{y}^{**}, \sigma_y^{2**})$, respectively.

(5) Evaluate test statistic: $t^{**} = \frac{\overline{x}^{**}-\overline{y}^{**}}{\sqrt{\sigma_x^{2**}/m+\sigma_y^{2**}/n}}$.

(6) Repeat steps 4 and 5 for B (e.g., 1000) times and obtain B values of the test statistic (t^{**}).

(7) Approximate *p*-value= $\frac{\text{number of times } (t^{**} \geq t_{obs})}{B}$.

The unpooled resampling method does not increase resampling variability because of small variety within each group. Therefore, it may reduce power of the test for small sample size studies. As such, a pooled resampling strategy in which bootstrap sample is drawn from pooled sample data obtained without centralizing each marginal sample (more appropriate for comparing distributions between groups) may perform better than the resampling strategy in which pooled sample data obtained after centralizing each marginal sample (more appropriate for comparing means). Thus, we used a pooled resampling scheme in which bootstrap sample is drawn from pooled sample data for conducting bootstrap tests, a similar approach suggested by Efron and Tibshirani, Algorithm 16.1, page 221 [32], for comparing equality of distributions. In short, the following steps were followed for computing the *p*-value using a nonparametric bootstrap test:

(1) Let $x = x_1, x_2 \ldots, x_m$ is the observed sample 1 of size m with mean $\bar{x}$ and $y = y_1, y_2 \ldots, y_n$ is the observed sample 2 of size m with mean $\bar{y}$.

(2) Evaluate test statistic: $t_{obs} = \frac{\overline{x}-\overline{y}}{\sqrt{\sigma_x^2/m+\sigma_y^2/n}}$

(3) Draw two bootstrap samples: one of size m observations with replacement (x^+) and another of size n observations with replacement (y^+) from the pooled sample. Compute mean and variance of each bootstrap sample as $(\overline{x}^{*+}, \sigma_x^{2*+})$ and $(\overline{y}^{*+}, \sigma_y^{2*+})$, respectively.

(4) Evaluate test statistic: $t^{*+} = \frac{\overline{x}^{*+}-\overline{y}^{*+}}{\sqrt{\sigma_x^{2*+}/m+\sigma_y^{2*+}/n}}$.

(5) Repeat steps 3 and 4 for B (e.g., 1000) times and obtain B values of the test statistic (t^{*+}).

(6) Approximate *p*-value=$\frac{\text{number of times } (t^{*+} \geq t_{obs})}{B}$.

The pooled resampling strategy described earlier would increase resampling variability and subsequently increase power of the test. In the paired group comparison, this pooled resampling strategy increases the variability; however, correlation between groups is neglected. Further, resampling data with replacement from the pooled data would increase more resampling variability than resampling data with replacement within each group separately under matched/paired design. A recent study referred this pooled approach as a counterintuitive resampling strategy for paired group comparison and showed asymptotically valid for comparing means [23]. In this study, the null hypothesis being tested is that the means of underlying distributions are the same, that is, H_0: $\mu_1 = \mu_2$ for comparison of two groups or H_0: $\mu 1 = \mu 2 = \mu 3$ for comparison of three groups, where μ_i is the mean for i^th^ group, i = (1,2,3). For Cauchy distributed data, the null hypothesis tested that the medians between two groups are same. However, in general, the proposed pooled bootstrap test, permutation test, and standard nonparametric tests examine whether the population distributions from which the samples were drawn are identical or not, H_0: F1 = F2 or H_0: F1 = F2 = F3 (where F_i is the marginal distribution for ith group, i = (1,2,3)), while parametric tests (unpaired *t*-test, paired *t*-test and *F*-test) directly compare whether the means of underlying distributions are same or not.

Comparison of parametric tests, standard nonparametric tests, permutation tests, and nonparametric bootstrap tests

The performance of parametric tests (unpaired *t*-test/Welch *t*-test, paired *t*-test, and one way analysis of variance test), corresponding exact nonparametric tests (Wilcoxon rank sum test, Wilcoxon signed-rank test, and Kruskal–Wallis test respectively), and corresponding nonparametric permutation tests and bootstrap tests (two independent means, two dependent means, and more than two independent means, respectively) under normal and non-normal distributions and various sample size conditions were evaluated. *P*-values for standard parametric and nonparametric tests can be easily computed using any standard statistical software. However, generally, we need to write our own function to compute nonparametric bootstrap *p*-value based on statistic of interest and method to compute *p*-value. The R codes are provided for computing bootstrap *p*-value using the aforementioned method for comparisons of two means, three means, and paired means (Appendix A).

Data generation for simulation studies

For comparing two or three independent means, first, we generated two or three datasets with considered means, standard deviations (SDs), and shapes. However, two correlated datasets with considered means, SDs, and shapes were generated for comparing two dependent means. Data were generated under condition (1) equal sample sizes and equal variances (2) equal sample sizes and unequal variances. These conditions were assessed under normal distribution and under skewed distributions (same skewness in groups and different skewness in groups). All the simulations were carried out for different sample sizes. The probability of type I error and empirical power were computed for each statistical test under different distributional assumptions and sample sizes. For type I error comparison, the original datasets were generated with same means with same or different SDs or skewness, while original datasets were generated with different means with same or different SDs or skewness for computing power of the test. After that, statistical tests were conducted to assess the significant difference between groups, and subsequently, *p*-values were recorded. These steps were repeated 10,000 times. Among converged solutions, the proportion of obtained significant *p*-value at 5% using each statistical test was estimated and compared.

R software was used for all simulations and analyses. The R command ‘rsnorm’ under the library ‘fGarch’ was used for generating normal or skew normal data for comparison of independent means. For generating skew normal data for paired situation, we used the multivariate skew normal distribution proposed by Azzalini and Dalla Valle [41] and used rmsn command under the library ‘sn’ for comparison of paired means (https://cran.r-project.org/web/packages/sn/sn.pdf) [42]. The robustness of these statistical tests was assessed by comparing type I error and statistical power across different tests.

Data examples

To demonstrate applications of our proposed nonparametric bootstrap test, we utilized datasets from two studies. The first example includes a partial dataset from an unpublished clinical trial study on epilepsy. The aim of this randomized controlled study was to compare the percent seizure reduction between an active arm and a control arm. In this study, we have compared percent change in seizure frequency between two groups among subjects who had a baseline seizure frequency greater than 18 seizures per month using an unpaired *t*-test, Welch *t*-test, nonparametric bootstrap *t*-test, exact Wilcoxon rank sum test, and asymptotic permutation *t*-test. The change in seizure frequency from baseline to post intervention among subjects who had baseline seizure frequency greater than 14 seizures per month was also compared using a paired *t*-test, nonparametric bootstrap paired *t*-test, exact Wilcoxon signed-rank test, and asymptotic permutation test for paired data separately for each treatment group. Further, the performance of the paired tests was assessed by random selection of data from each treatment group with different sample sizes. The second example, a published and publically available multisite study, is a randomized clinical trial on motivational interviewing to improve treatment engagement and outcome in the subjects seeking treatment for substance use as compared with a standard intervention. The primary outcomes were treatment retention and substance use at 28 days and at 84 days after randomization to motivational intervention (MI) and standard intervention (SI) [43]. In the present study, the number of treatment retention at 28 days was compared using unpaired *t*-test, Welch *t*-test, nonparametric bootstrap *t*-test, exact Wilcoxon rank sum test, and asymptotic permutation *t*-test. The performance of these tests was also assessed by random selection of data from each treatment group with different sample sizes. The sample size is denoted using N, and data are summarized using mean and SD.

Results

Table I summarizes the minimum sample size needed to conduct exact nonparametric tests. This table clearly displays that the exact Wilcoxon rank sum test can be used only if sample size is at least eight, while the exact Wilcoxon signed-rank test can be used only if the sample size is greater than or equal to

Table I. Minimum sample sizes for using common nonparametric exact tests.

	Wilcoxon rank sum test	Wilcoxon signed-rank test	Kruskal–Wallis test with 3 groups	Friedman test with 3 groups
Equal groups	$N = 8$ (k = 2, $n = 4$)	$N = 6$ (k = 2, $n = 3$)	$N = 9$ (k = 3, $n = 3$)	$N = 9$ (k = 3, $n = 3$)
Unequal groups	$N = 8$ (k = 2, $n_1 = 5$ $n_2 = 3$)	Not applicable	$N = 8$ (k = 3, $n_1 = 4$ $n_2 = 2$ $n_3 = 2$) or (k = 3, $n_1 = 2$ $n_2 = 3$ $n_3 = 3$)	Not applicable

k = number of groups, n = sample size within groups; N = total sample size.

six. When using asymptotic nonparametric tests, a sample size of at least 16 is required for using Wilcoxon rank and signed-rank tests, while 24 observations are needed for asymptotic Kruskal–Wallis test with four groups [5].

Table II shows the type I error probability for Student's *t*-test, Welch *t*-test, exact Wilcoxon rank sum test, asymptotic permutation test, and nonparametric bootstrap *t*-test. All the tests reasonably controlled type I error probability when variances were equal between groups, irrespective of normal or skew normal distributions. However, when variances were not equal between groups, then all tests yielded slightly high type I error probabilities up to 11%. Nonparametric bootstrap *t*-test produced slightly high type I error (range 5–8%) as opposed to other tests for unequal sample sizes with equal variances condition. In case of unequal sample sizes between groups with unequal variances, unpaired *t*-test (range 6–12%), Wilcoxon rank sum test (range 0–12%), nonparametric bootstrap test (range 6–20%), and asymptotic permutation *t*-test (4–16%) produced very high type I error probabilities when the total sample size was set up below 10. The type I error probabilities of the Student's *t*-test, Welch *t*-test, exact Wilcoxon rank sum test, and asymptotic permutation *t*-test were close to a nominal level of 5%, whereas the type I error probabilities for nonparametric bootstrap *t*-test were found to be between 5% and 7% when the sample size was set up greater than or equal to 10.

The power comparison of Student's *t*-test, Welch *t*-test, exact Wilcoxon rank sum test, asymptotic permutation *t*-test, and nonparametric bootstrap *t*-test under a variety of situations are displayed in Table III. In all the conditions, nonparametric bootstrap *t*-test yielded greater or equal power compared with unpaired *t*-test, asymptotic permutation *t*-test, Welch *t*-test, and exact Wilcoxon rank sum test. Generally, nonparametric pooled bootstrap *t*-test produced larger power compared with unpaired *t*-test followed by asymptotic permutation *t*-test or Welch *t*-test and exact Wilcoxon rank sum test. In all conditions, except unequal sample sizes and unequal variances, Wilcoxon rank sum test produced the least amount of power as compared with the other tests, followed by the Welch *t*-test or permutation *t*-test. The permutation *t*-test produced larger power compared with Welch *t*-test for unequal variances and equal sample sizes otherwise slightly lower or equal to Welch *t*-test. Approximately, 1–4% gain in power was found with nonparametric bootstrap *t*-test as compared with unpaired *t*-test. However, a marked gain in power was observed for nonparametric bootstrap *t*-test as compared with unpaired *t*-test for unequal sample sizes and unequal variances and when the sample size was greater than or equal to 10. No test provided valid *p*-values for unequal sample sizes and unequal variances when the sample size was below or equal to nine. For unequal sample sizes and unequal variances with skew normal data, nonparametric bootstrap *t*-test provided greater power followed by Welch *t*-test, Wilcoxon rank sum test, Student's *t*-test, and asymptotic permutation *t*-test for sample size greater than or equal to 10.

All the statistical tests for comparing two independent means yielded type I error close to nominal level or slightly higher (6%) for all distributions (lognormal, Poisson, and chi square; Table S1). Table IV shows the power comparison of conventional statistical tests with nonparametric bootstrap *t*-test and asymptotic and exact methods of permutation *t*-test for comparison of two independent groups under different distributions. The pooled bootstrap *t*-test demonstrated power advantage compared with other statistical tests for all distributions except for heavy-tailed distributions (lognormal and Cauchy). The pooled bootstrap *t*-test produced a marked improvement in power compared with other alternatives especially in case of unequal group sizes. For extreme variable lognormal and Cauchy distributions, exact Wilcoxon rank sum test yielded greater power followed by exact permutation *t*-test and pooled nonparametric bootstrap *t*-test. Generally, asymptotic permutation *t*-test produced equal or slightly more power compared with exact permutation *t*-test except for lognormal and Cauchy distributions.

Tables V and VI depict type I error probability and power for statistical tests to compare paired/matched means respectively. For normal distribution with equal and unequal variances, all four statistical tests

Table II. The type I error probability (nominal level = 0.05) for Student's *t*-test (ST), Welch *t*-test (WT), nonparametric bootstrap *t*-test (NPBTT), exact Wilcoxon rank sum test (WRST), and permutation *t*-test (PTT) under various conditions for different sample sizes.

Sample size (per group)	Normal and equal variance F1 = N(M = 5, S = 1) vs. F2 = N(M = 5, S = 1)					Normal and unequal variance F1 = N(M = 5, S = 1) vs. F2 = N(M = 5, S = 3)				
	ST	WT	NPBTT	WRST	PTT	ST	WT	NPBTT	WRST	PTT
3	0.05	0.04	0.06	0.00	0.02	0.07	0.05	0.09	0.00	0.04
4	0.05	0.04	0.06	0.03	0.04	0.07	0.05	0.08	0.05	0.05
5	0.05	0.05	0.05	0.03	0.04	0.07	0.05	0.07	0.05	0.06
6	0.05	0.05	0.05	0.04	0.04	0.06	0.05	0.07	0.05	0.06
7	0.05	0.05	0.05	0.04	0.05	0.06	0.05	0.07	0.05	0.06
8	0.05	0.05	0.05	0.05	0.05	0.06	0.05	0.07	0.07	0.06
9	0.05	0.05	0.05	0.04	0.05	0.06	0.05	0.06	0.06	0.05
10	0.05	0.05	0.05	0.04	0.05	0.05	0.05	0.05	0.06	0.05
15	0.04	0.04	0.04	0.04	0.04	0.05	0.05	0.06	0.06	0.05
	Same skewed and equal variance F1 = SN(M = 5, S = 1, Sk = 0.8) vs. F2 = SN(M = 5, S = 1, Sk = 0.8)					Same skewed and unequal variance F1 = SN(M = 5, S = 1, Sk = 0.8) vs. F2 = SN(M = 5, S = 3 Sk = 0.8)				
3	0.05	0.03	0.06	0.00	0.02	0.09	0.07	0.10	0.00	0.05
4	0.05	0.04	0.05	0.03	0.04	0.09	0.07	0.09	0.06	0.07
5	0.05	0.04	0.05	0.03	0.04	0.07	0.06	0.08	0.05	0.06
6	0.05	0.04	0.05	0.04	0.04	0.08	0.07	0.08	0.05	0.07
7	0.05	0.04	0.05	0.04	0.04	0.07	0.06	0.07	0.07	0.06
8	0.05	0.05	0.05	0.05	0.05	0.07	0.06	0.07	0.08	0.07
9	0.05	0.05	0.05	0.04	0.05	0.07	0.06	0.07	0.07	0.06
10	0.05	0.05	0.05	0.04	0.05	0.07	0.06	0.07	0.08	0.06
15	0.05	0.05	0.05	0.04	0.05	0.06	0.05	0.06	0.09	0.06
	Different skewed and equal variance F1 = SN(M = 5, S = 1, Sk = 0.8) vs. F2 = SN(M = 5, S = 1, Sk = 1.0)					Different skewed and unequal variance F1 = SN(M = 5, S = 1, Sk = 0.8) vs. F2 = SN(M = 5, S = 3, Sk = 1.0)				
3	0.05	0.04	0.06	0.00	0.02	0.10	0.07	0.11	0.00	0.05
4	0.05	0.04	0.05	0.03	0.04	0.09	0.08	0.10	0.06	0.08
5	0.04	0.04	0.05	0.03	0.04	0.08	0.07	0.08	0.05	0.07
6	0.05	0.04	0.05	0.04	0.04	0.08	0.07	0.08	0.06	0.07
7	0.05	0.04	0.05	0.04	0.04	0.07	0.06	0.07	0.07	0.07
8	0.05	0.04	0.05	0.05	0.04	0.07	0.07	0.08	0.09	0.07
9	0.05	0.05	0.05	0.04	0.05	0.07	0.07	0.08	0.08	0.07
10	0.05	0.05	0.05	0.04	0.04	0.07	0.06	0.07	0.09	0.06
15	0.05	0.05	0.05	0.05	0.05	0.07	0.06	0.07	0.10	0.06
	Unequal sample size, same skewed, and equal variance F1 = SN(M = 5, S = 1, Sk = 0.8) vs. F2 = SN(M = 5, S = 1, Sk = 0.8)					Unequal sample size, same skewed, and unequal variance F1 = SN(M = 5, S = 1, Sk = 0.8) vs. F2 = SN(M = 5, S = 3, Sk = 0.8)				
4, 2	0.05	0.05	0.08	0.00	0.02	0.20	0.12	0.20	0.00	0.12
3, 4	0.05	0.04	0.06	0.00	0.03	0.06	0.06	0.08	0.00	0.04
5, 3	0.05	0.05	0.06	0.03	0.03	0.16	0.09	0.14	0.12	0.13
4, 5	0.05	0.04	0.05	0.03	0.04	0.06	0.06	0.08	0.04	0.05
6, 3	0.05	0.05	0.07	0.05	0.04	0.18	0.09	0.15	0.15	0.16
4, 6	0.05	0.05	0.06	0.04	0.04	0.05	0.06	0.07	0.04	0.04
3, 7	0.05	0.06	0.08	0.03	0.04	0.02	0.05	0.06	0.02	0.02
4, 11	0.05	0.06	0.08	0.04	0.04	0.01	0.05	0.05	0.02	0.01
5, 10	0.05	0.05	0.06	0.04	0.04	0.02	0.05	0.05	0.03	0.02
6, 9	0.04	0.04	0.05	0.05	0.04	0.04	0.06	0.06	0.05	0.03

Simulation parameters under normal (N) and skewed normal (SN) distributions with given mean (M), standard deviation (S), and Skewness (Sk).

retained nominal false positive probability (i.e., 0.05). In all conditions, asymptotic permutation paired *t*-test reasonably controlled type I error. Maximum type I error probability for paired *t*-test and exact Wilcoxon signed-rank test was found to be 6% while 7% for nonparametric bootstrap paired *t*-test.

Table III. Statistical power for Student's *t*-test (ST), Welch *t*-test (WT), nonparametric bootstrap *t*-test (NPBTT), exact Wilcoxon rank sum test (WRST), and permutation *t*-test (PTT) under various conditions for different sample sizes.

Sample size (per group)	Normal and equal variance F1 = N(M = 5, S = 1) vs. F2 = N(M = 5, S = 1)					Normal and unequal variance F1 = N(M = 5, S = 1) vs. F2 = N(M = 5, S = 3)				
	ST	WT	NPBTT	WRST	PTT	ST	WT	NPBTT	WT	PTT
3	0.46	0.36	0.49	0.00	0.27	0.17	0.12	0.20	0.00	0.10
4	0.65	0.59	0.66	0.48	0.57	0.22	0.17	0.24	0.17	0.18
5	0.79	0.76	0.79	0.68	0.75	0.26	0.21	0.28	0.20	0.23
6	0.88	0.86	0.88	0.82	0.86	0.31	0.26	0.32	0.24	0.28
7	0.93	0.92	0.93	0.88	0.92	0.36	0.31	0.37	0.31	0.34
8	0.96	0.95	0.96	0.95	0.95	0.40	0.36	0.41	0.38	0.39
9	0.98	0.98	0.98	0.96	0.98	0.44	0.40	0.45	0.37	0.42
10	0.99	0.99	0.99	0.98	0.99	0.48	0.45	0.49	0.43	0.47
15	1.00	1.00	1.00	1.00	1.00	0.66	0.64	0.66	0.61	0.65
25	1.00	1.00	1.00	1.00	1.00	0.88	0.87	0.88	0.84	0.87
50	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.98	0.99
100	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Same skewed and equal variance F1 = SN(M = 5, S = 1, Sk = 0.8) vs. F2 = SN(M = 5, S = 1, Sk = 0.8)					Same skewed and unequal variance F1 = SN(M = 5, S = 1, Sk = 0.8) vs. F2 = SN(M = 5, S = 3, Sk = 0.8)				
3	0.48	0.38	0.51	0.00	0.29	0.12	0.08	0.15	0.00	0.06
4	0.67	0.62	0.68	0.51	0.60	0.16	0.11	0.18	0.13	0.13
5	0.80	0.78	0.80	0.68	0.77	0.21	0.15	0.23	0.16	0.18
6	0.87	0.86	0.87	0.82	0.86	0.26	0.20	0.28	0.20	0.23
7	0.92	0.92	0.92	0.88	0.92	0.31	0.25	0.33	0.24	0.29
8	0.95	0.95	0.95	0.94	0.95	0.36	0.31	0.38	0.30	0.34
9	0.97	0.97	0.97	0.96	0.97	0.40	0.35	0.42	0.29	0.38
10	0.99	0.99	0.99	0.98	0.99	0.45	0.41	0.47	0.33	0.44
15	1.00	1.00	1.00	1.00	1.00	0.67	0.65	0.69	0.50	0.67
25	1.00	1.00	1.00	1.00	1.00	0.91	0.90	0.91	0.73	0.91
50	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.95	1.00
100	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Different skewed and equal variance F1 = SN(M = 5, S = 1, Sk = 0.8) vs. F2 = SN(M = 5, S = 1, Sk = 1.0)					Different skewed and unequal variance F1 = SN(M = 5, S = 1, Sk = 0.8) vs. F2 = SN(M = 5, S = 3, Sk = 1.0)				
3	0.48	0.38	0.52	0.00	0.29	0.12	0.07	0.14	0.00	0.06
4	0.67	0.63	0.69	0.53	0.60	0.15	0.10	0.17	0.11	0.11
5	0.80	0.78	0.80	0.68	0.77	0.19	0.13	0.22	0.15	0.16
6	0.88	0.87	0.88	0.83	0.86	0.25	0.19	0.28	0.19	0.22
7	0.93	0.92	0.93	0.89	0.92	0.29	0.24	0.32	0.21	0.27
8	0.96	0.96	0.96	0.95	0.95	0.36	0.30	0.38	0.28	0.34
9	0.98	0.98	0.98	0.96	0.98	0.40	0.35	0.43	0.27	0.38
10	0.99	0.99	0.99	0.98	0.99	0.45	0.41	0.48	0.31	0.44
15	1.00	1.00	1.00	1.00	1.00	0.68	0.66	0.70	0.45	0.67
25	1.00	1.00	1.00	1.00	1.00	0.92	0.91	0.92	0.68	0.91
50	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.94	1.00
100	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Unequal sample size, same skewed, and equal variance F1 = SN(M = 5, S = 1, Sk = 0.8) vs. F2 = SN(M = 5, S = 1, Sk = 0.8)					Unequal sample size, same skewed, and unequal variance F1 = SN(M = 5, S = 1, Sk = 0.8) vs. F2 = SN(M = 5, S = 3, Sk = 0.8)				
4, 2	0.44	0.30	0.49	0.00	0.27	0.24	0.11	0.22	0.00	0.16
3, 4	0.58	0.52	0.61	0.00	0.46	0.10	0.10	0.15	0.00	0.06
5, 3	0.64	0.52	0.66	0.50	0.57	0.25	0.09	0.19	0.18	0.21
4, 5	0.74	0.69	0.74	0.62	0.69	0.14	0.15	0.21	0.13	0.11
6, 3	0.68	0.54	0.71	0.61	0.63	0.29	0.10	0.20	0.24	0.26
4, 6	0.78	0.72	0.78	0.69	0.75	0.14	0.20	0.26	0.14	0.11

(*Continues*)

Table III. *(Continued)*

Sample size (per group)	Normal and equal variance F1 = N(M = 5, S = 1) vs. F2 = N(M = 5, S = 1)					Normal and unequal variance F1 = N(M = 5, S = 1) vs. F2 = N(M = 5, S = 3)				
	ST	WT	NPBTT	WRST	PTT	ST	WT	NPBTT	WT	PTT
3, 7	0.74	0.60	0.74	0.64	0.70	0.07	0.24	0.28	0.09	0.06
4, 11	0.89	0.75	0.85	0.79	0.88	0.10	0.42	0.45	0.14	0.09
5, 10	0.92	0.85	0.89	0.86	0.91	0.17	0.39	0.43	0.19	0.15
6, 9	0.93	0.90	0.92	0.91	0.93	0.24	0.36	0.41	0.25	0.22

Simulation parameters under normal (N) and skewed normal (SN) distributions with given mean (M), standard deviation (S), and Skewness (Sk).

Table IV. Statistical power for Student's *t*-test (ST), Welch *t*-test (WT), nonparametric bootstrap *t*-test (NPBTT), exact Wilcoxon rank sum test (WRST), asymptotic permutation *t*-test (PTT$_A$), and exact permutation *t*-test (PTT$_E$) under different distributions for different sample sizes.

N_1	N_2	ST	WT	NPBTT	WRST	PTT$_A$	PTT$_E$
LN (1, 0.6)	LN (2, 1)						
5	5	0.23	0.14	0.28	0.31	0.19	0.29
5	10	0.17	0.51	0.58	0.45	0.15	0.20
10	10	0.60	0.53	0.64	0.68	0.58	0.76
P (5)	P (10)						
5	5	0.72	0.68	0.76	0.63	0.68	0.65
5	10	0.86	0.88	0.91	0.84	0.85	0.84
10	10	0.98	0.97	0.98	0.96	0.97	0.97
$\chi^2(3)$	$\chi^2(6)$						
5	5	0.30	0.26	0.34	0.26	0.27	0.27
5	10	0.37	0.50	0.54	0.41	0.33	0.37
10	10	0.58	0.57	0.61	0.60	0.57	0.58
LN (1, 0.6)	LN (3, 4)						
5	5	0.03	0.01	0.04	0.15	0.02	0.06
5	10	0.00	0.03	0.06	0.15	0.00	0.00
10	10	0.06	0.04	0.07	0.33	0.05	0.20
Cauchy (5, 2)	Cauchy (10, 4)						
5	5	0.13	0.11	0.15	0.17	0.11	0.15
5	10	0.10	0.13	0.15	0.24	0.09	0.11
10	10	0.14	0.13	0.16	0.37	0.14	0.20
$\chi^2(6)$	P (10)						
5	5	0.44	0.41	0.47	0.34	0.40	0.36
5	10	0.58	0.53	0.59	0.51	0.56	0.55
10	10	0.72	0.72	0.74	0.73	0.71	0.70
LN (1, 0.6)	$\chi^2(6)$						
5	5	0.27	0.22	0.31	0.22	0.23	0.24
5	10	0.31	0.49	0.52	0.35	0.28	0.30
10	10	0.55	0.53	0.58	0.54	0.54	0.55

N_1, sample size for group 1; N_2, sample size for group 2; LN, lognormal; P, Poisson.

The nonparametric bootstrap paired *t*-test provided the highest power compared with paired *t*-test, asymptotic permutation paired *t*-test, and exact Wilcoxon signed-rank test. The gain in power with nonparametric bootstrap paired *t*-test ranged from 1% to 8% compared with paired *t*-test. Exact Wilcoxon signed-rank test demonstrated minimum power compared with other tests. The percentage of bootstrap sample with SD equal to 0 in both groups was less than 0.1% for any paired or unpaired conditions.

Type I error probability and power for statistical tests to compare three means were evaluated under different conditions. For equal variances irrespective of normal or skewed distributions, all the tests produced a nominal level of the type I error probability. However, *F*-test and nonparametric bootstrap *F*-test produced type I error probability as high as 9%, which is substantially greater than the acceptable level

Table V. The type I error probability (nominal level = 0.05) for paired *t*-test (PT), nonparametric bootstrap paired *t*-test (NPBPT), exact Wilcoxon signed-rank test (WSRT), and permutation paired *t*-test (PPTT) under various conditions for different sample sizes.

Sample size	Normal and equal variance F1 = N(M = 5, S = 1) vs. F2 = N(M = 5, S = 1)				Normal and unequal variance F1 = N(M = 5, S = 1) vs. F2 = N(M = 5, S = 3)				Same skewed and equal variance F1 = SN(M = 5, S = 1, Sk = 0.5) vs. F2 = SN(M = 5, S = 1, Sk = 0.5)			
	PT	NPBPT	WSRT	PPTT	PT	NPBPT	WSRT	PPTT	PT	NPBPT	WSRT	PPTT
3	0.05	0.07	0.00	0.00	0.05	0.07	0.00	0.00	0.05	0.07	0.00	0.00
4	0.05	0.07	0.00	0.00	0.05	0.07	0.00	0.00	0.05	0.07	0.00	0.00
5	0.05	0.06	0.00	0.02	0.05	0.06	0.00	0.02	0.05	0.05	0.00	0.02
6	0.05	0.05	0.03	0.03	0.05	0.05	0.03	0.03	0.05	0.05	0.03	0.03
7	0.05	0.05	0.05	0.04	0.05	0.05	0.05	0.04	0.05	0.05	0.05	0.04
8	0.05	0.05	0.04	0.04	0.05	0.05	0.04	0.04	0.05	0.05	0.04	0.04
9	0.05	0.05	0.04	0.04	0.05	0.05	0.04	0.04	0.05	0.05	0.04	0.04
10	0.05	0.05	0.05	0.04	0.05	0.05	0.05	0.04	0.05	0.05	0.05	0.04
15	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
25	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
	Same skewed and unequal variance F1 = SN(M = 5, S = 1, Sk = 0.5) vs. F2 = SN(M = 5, S = 3, Sk = 0.5)				Different skewed and equal variance F1 = SN(M = 5, S = 1, Sk = 0.2) vs. F2 = SN(M = 5, S = 1, Sk = 0.8)				Different skewed and unequal variance F1 = SN(M = 5, S = 1, Sk = 0.2) vs. F2 = SN(M = 5, S = 3, Sk = 0.8)			
3	0.05	0.07	0.00	0.00	0.05	0.07	0.00	0.00	0.05	0.07	0.00	0.00
4	0.05	0.07	0.00	0.00	0.05	0.07	0.00	0.00	0.05	0.07	0.00	0.00
5	0.05	0.06	0.00	0.02	0.05	0.05	0.00	0.02	0.05	0.05	0.00	0.02
6	0.05	0.05	0.03	0.03	0.05	0.05	0.03	0.03	0.05	0.05	0.03	0.03
7	0.05	0.05	0.05	0.04	0.05	0.05	0.05	0.04	0.05	0.05	0.05	0.04
8	0.05	0.05	0.04	0.04	0.05	0.05	0.04	0.04	0.05	0.05	0.04	0.04
9	0.06	0.06	0.04	0.05	0.05	0.05	0.04	0.04	0.05	0.05	0.04	0.04
10	0.05	0.05	0.05	0.04	0.05	0.05	0.05	0.04	0.05	0.05	0.05	0.04
15	0.06	0.06	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
25	0.06	0.06	0.06	0.06	0.06	0.05	0.05	0.05	0.05	0.05	0.05	0.05

Simulation parameters under normal (N) and skewed normal (SN) distributions with given mean (M), standard deviation (S), Skewness (Sk), and correlation = 0.80.

Table VI. Statistical power for paired *t*-test (PT), nonparametric bootstrap paired *t*-test (NPBPT), exact Wilcoxon signed-rank test (WSRT), and permutation paired *t*-test (PPTT) under various conditions for different sample sizes.

Sample size	Normal and equal variance F1 = N(M = 5, S = 1) vs. F2 = N(M = 5, S = 1)				Normal and unequal variance F1 = N(M = 5, S = 1) vs. F2 = N(M = 5, S = 3)				Same skewed and equal variance F1 = SN(M = 5, S = 1, Sk = 0.5) vs. F2 = SN(M = 5, S = 1, Sk = 0.5)			
	PT	NPBPT	WSRT	PPTT	PT	NPBPT	WSRT	PPTT	PT	NPBPT	WSRT	PPTT
3	0.78	0.85	0.00	0.00	0.26	0.33	0.00	0.00	0.78	0.85	0.00	0.00
4	0.98	0.99	0.00	0.35	0.43	0.51	0.00	0.05	0.98	0.99	0.00	0.35
5	1.00	1.00	0.00	0.98	0.59	0.63	0.00	0.38	1.00	1.00	0.00	0.99
6	1.00	1.00	1.00	1.00	0.72	0.73	0.54	0.60	1.00	1.00	1.00	1.00
7	1.00	1.00	1.00	1.00	0.81	0.82	0.78	0.75	1.00	1.00	1.00	1.00
8	1.00	1.00	1.00	1.00	0.88	0.88	0.82	0.84	1.00	1.00	1.00	1.00
9	1.00	1.00	1.00	1.00	0.92	0.92	0.88	0.90	1.00	1.00	1.00	1.00
10	1.00	1.00	1.00	1.00	0.95	0.95	0.94	0.94	1.00	1.00	1.00	1.00
15	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00
25	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
50	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

(*Continues*)

Table VI. *(Continued)*

Sample size	Normal and equal variance F1 = N(M = 5, S = 1) vs. F2 = N(M = 5, S = 1)				Normal and unequal variance F1 = N(M = 5, S = 1) vs. F2 = N(M = 5, S = 3)				Same skewed and equal variance F1 = SN(M = 5, S = 1, Sk = 0.5) vs. F2 = SN(M = 5, S = 1, Sk = 0.5)			
	PT	NPBPT	WSRT	PPTT	PT	NPBPT	WSRT	PPTT	PT	NPBPT	WSRT	PPTT
100	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Same skewed and unequal variance F1 = SN(M = 5, S = 1, Sk = 0.5) vs. F2 = SN(M = 5, S = 3, Sk = 0.5)				Different skewed and equal variance F1 = SN(M = 5, S = 1, Sk = 0.2) vs. F2 = SN(M = 5, S = 1, Sk = 0.8)				Different skewed and unequal variance F1 = SN(M = 5, S = 1, Sk = 0.2) vs. F2 = SN(M = 5, S = 3, Sk = 0.8)			
3	0.30	0.37	0.00	0.00	0.79	0.85	0.00	0.00	0.42	0.51	0.00	0.00
4	0.49	0.57	0.00	0.06	0.98	0.99	0.00	0.36	0.70	0.77	0.00	0.11
5	0.67	0.71	0.00	0.45	1.00	1.00	0.00	0.99	0.87	0.88	0.00	0.69
6	0.81	0.82	0.62	0.70	1.00	1.00	1.00	1.00	0.95	0.96	0.83	0.90
7	0.89	0.89	0.86	0.84	1.00	1.00	1.00	1.00	0.98	0.98	0.97	0.97
8	0.93	0.94	0.90	0.91	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99
9	0.96	0.96	0.93	0.95	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00
10	0.98	0.98	0.97	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
15	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
25	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
50	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
100	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Simulation parameters under normal (N) and skewed normal distributions (SN) with given mean (M), standard deviation (S), Skewness (Sk), and correlation = 0.80.

of 5%, whereas Kruskal–Wallis test and asymptotic permutation k-sample test yielded type I error probability up to 7% for normal distributions with unequal variances. For different skewed normal distributions with unequal variances, all tests yielded high type I error probabilities close to 10%. F-test and nonparametric bootstrap F-test produced very similar statistical power, greater than asymptotic permutation k-sample test and Kruskal–Wallis test in all the conditions. Under equal variances, asymptotic permutation k-sample test produced greater power compared with Kruskal–Wallis test, while Kruskal–Wallis test demonstrated power advantage as compared with permutation test when variances were unequal. As expected, statistical power increased with an increase in sample size. As anticipated, the differences in statistical power were reduced with an increase in sample size (Tables S2 and S3).

In the first data example, the mean percent change in seizures was observed as 0.56 (SD: 0.10) and 0.19 (SD: 0.11) in active arm (n = 6) and control (n = 5) arm, respectively, among subjects who had more than 18 seizures per month at baseline. Nonparametric bootstrap t-test (p = 0.032), Student's t-test (p = 0.039), Welch t-test (p = 0.04), and permutation t-test (p = 0.048) showed a significant difference between groups, while the exact Wilcoxon rank sum test (p = 0.08) showed a nonsignificant p-value. The analysis of change in seizure frequency between pre-intervention and post-intervention among patients who had baseline seizure frequency greater than 14 for active group (n = 10) showed mean seizure frequency as 24.05 (SD: 8.15) and 11.87 (SD: 6.94) for pre-intervention and post-intervention, respectively. The reduction in seizure frequency was found to be statistically significant with all tests (p = 0.0007 using paired t-test, p = 0.001 using nonparametric bootstrap paired t-test, p = 0.004 using Wilcoxon signed-rank test, and p = 0.006 with permutation paired t-test). In the control group (n = 7), the mean seizure frequency at baseline and post intervention was 38.89 (SD: 32.89) and 30.58 (SD: 28.52), respectively. This change was found statistically significant using nonparametric bootstrap paired t-test paired t-test (p = 0.044), while paired t-test (p = 0.06), permutation test (p = 0.07), and Wilcoxon signed-rank test (p = 0.09) demonstrated no statistically significant difference between groups. The comparison of these tests is displayed for different sample sizes in Table VII. It showed that the paired t-test, asymptotic permutation paired t-test, and nonparametric bootstrap paired t-test provided similar p-values for mild skewed data. However, nonparametric bootstrap paired t-test and Wilcoxon signed-rank test provided similar and lower p-values for extreme variable data compared with paired t-test and permutation paired t-test. In the second example, the mean number of retention

Table VII. Comparisons of *p*-values obtained using paired *t*-test (PT), nonparametric bootstrap paired *t*-test (NPBPT), exact Wilcoxon signed-rank test (WSRT), and permutation paired *t*-test (PPTT) for comparing pre-intervention and post-intervention seizure frequencies in the real dataset.

Group	Sample size	Pre-seizures		Post-seizures		Correlation	*p*-value			
		Mean	SD	Mean	SD		PT	NPBPT	WSRT	PPTT
Active	20	12.08	10.35	8.20	6.06	0.54	0.06	0.06	0.10	0.06
	10	11.15	8.41	9.63	5.62	0.84	0.34	0.35	0.64	0.31
	6	15.25	8.49	12.22	5.87	0.77	0.23	0.23	0.44	0.20
Control	20	16.40	24.83	12.77	20.34	0.98	0.02	0.01	0.01	0.03
	10	25.10	32.92	19.63	27.32	0.98	0.07	0.06	0.04	0.07
	6	20.33	25.13	16.33	25.25	0.99	0.07	0.03	0.03	0.08

SD, standard deviation

was obtained as 5.02 (SD: 5.11) in the MI group and 4.07 (SD: 4.15) in the SI group. The distribution and comparison of number of retention at 28 days obtained through random selections of data from each treatment group separately with different sample sizes are provided in Table VIII. The *p*-values obtained from different statistical tests except Wilcoxon rank sum test were found to be similar for equal sample sizes. Table VIII clearly demonstrated that the obtained *p*-values using Welch *t*-test and nonparametric bootstrap *t*-test were very similar but different than other tests for unequal sample sizes and unequal variances.

Discussion

Small sample size studies are very common in basic and health science research [16]. The feasibility and validity of nonparametric bootstrap test with a pooled resampling method compared with standard parametric, nonparametric, and permutation tests have not been studied extensively for small sample size studies. In simulation studies, we found that nonparametric bootstrap *t*-test and nonparametric bootstrap paired *t*-test were equal or more powerful than unpaired and paired *t*-tests, respectively, in all situations. The performance of bootstrap test was also found similar to one way analysis of variance test. Nonparametric bootstrap test was found always superior to their alternative standard nonparametric or permutation tests except for a few occasions. In most of the simulated cases for nonparametric bootstrap tests, the type I error probabilities did not exceed much from the nominal level of significance.

Kang and Harring have compared the performance of nonparametric bootstrap test with Wilcoxon rank sum test and Welch *t*-test [22]. Similar to our study, this study found that the nonparametric bootstrap test has better power as compared with the unpaired *t*-test, Welch *t*-test, and Wilcoxon rank

Table VIII. Comparisons of *p*-values obtained using Student's *t*-test (ST), Welch *t*-test (WT), nonparametric bootstrap *t*-test (NPBTT), exact Wilcoxon rank sum test (WRST), and permutation *t*-test (PTT) for comparing two treatment groups in the real dataset.

MI			SI			*p*-value				
N	Mean	SD	N	Mean	SD	ST	WT	NPBTT	WRST	PTT
173	5.02	5.11	177	4.07	4.15	0.06	0.06	0.06	0.15	0.06
50	5.60	5.75	50	4.30	4.38	0.21	0.21	0.19	0.43	0.20
50	5.60	5.75	20	3.80	3.53	0.20	0.12	0.11	0.44	0.20
50	5.60	5.75	10	2.90	2.60	0.15	0.03	0.02	0.25	0.15
20	5.25	5.91	20	4.25	3.21	0.51	0.51	0.52	0.93	0.50
20	5.25	5.91	10	3.90	3.00	0.50	0.41	0.42	0.95	0.50
20	5.25	5.91	5	2.00	2.00	0.24	0.05	0.04	0.32	0.24
10	6.10	6.40	10	3.40	3.20	0.25	0.25	0.26	0.36	0.24
10	6.10	6.40	5	4.00	3.67	0.51	0.44	0.46	0.71	0.49
6	4.00	3.63	3	11.00	9.54	0.14	0.33	0.25	0.24	0.13

MI, motivational intervention; SI, standard intervention; SD, standard deviation.

sum test especially in the situation where the two groups showed different patterns of non-normality. In this study, authors have used mean difference in bootstrap test as a statistic of interest as opposed to our study. Further, this study did not test the performance of bootstrap test in extremely small sample size studies ($N < 16$). Ahad *et al.* examined the relative power performance of unpaired t-test compared with bootstrap test under unequal group sizes ($N = 20$) [25]. Like our study, this study concluded that the bootstrap test outperformed the pooled t-test under all assessed conditions. Kulkarni evaluated the type I error probabilities of unpooled and pooled nonparametric bootstrap test as compared with Student's t-test for comparing two independent means using extensive simulations [21]. This study revealed that unpooled error nonparametric bootstrap test outperformed t-test in case of heterogeneous variance and unequal sample sizes, and when sample size is moderately large. In very small sample size studies (<14), both bootstrap methods showed unacceptably high type I error probabilities. However, this study did not evaluate power comparison across different tests. Lansing extended the work of Kurlkarni by determining power comparison of unpooled and pooled nonparametric bootstrap test with Students' t-test and Welch t-test [20]. In contrast to our study, this study concluded that bootstrap procedure is not feasible for small sample size studies and provides as powerful as t-test when sample sizes are large. These studies used a different bootstrap method compared with our study.

In another study [24], permutation test, studentized bootstrap t-test, Welch t-test, and Wilcoxon rank sum test were compared under heteroscedastic distributions for small sample size studies. Similar to our findings, this study found that studentized permutation test outperforms Welch t-test for unequal variances and equal sample sizes. In unequal sample sizes and skewed distributions, this study noticed unsatisfactory results. We further demonstrated that permutation t-test produced lower or equal power compared with Welch t-test for unequal sample sizes. In concordance with our findings, this study also observed a higher statistical power for studentized bootstrap t-test compared with Welch t-test. However, this study did not evaluate the performance of studentized bootstrap t-test for extremely small sample size situations ($N < 16$) and using a pooled resampling method. We observed marked gain in power with proposed bootstrap t-test compared with permutation t-test. To support this, Janssen and Pauls [24] also mentioned that bootstrap test may be superior for exponential densities with unequal sample sizes, unequal variances, and skewed distributions. In our study, we did not find benefit of using proposed bootstrap t-test compared with permutation and Wilcoxon rank sum test for heavy tails distribution (Cauchy distribution). Based on theoretical and applied works, Janssen and Pauls [24,44] also recommended that bootstrap test should not be used for Cauchy distributed data. For heavy-tailed distributions like Cauchy distribution, we suggest using exact Wilcoxon rank sum test as opposed to permutation test. We found that asymptotic permutation t-test generally yielded more powerful results compared with exact permutation t-test. Similarly, Corcoran and Mehta [36] showed that asymptotic permutation test of trend produces uniformly larger power than exact permutation test of trend. In most cases, the power of the permutation t-test was found to be higher compared with exact Wilcoxon rank sum test in current study. Similar to our study, a number of other studies described power advantage with permutation test compared with Wilcoxon rank sum test [24,37,45]. In contrary, Wilcoxon rank sum test yielded remarkably powerful results than both t-test and permutation test in a study conducted by Weber and Sawilowsky [9].

In another study [39], authors compared performance of nonparametric bootstrap test using a different test statistic with Student's t-test, Welch t-test, and Wilcoxon rank sum test under unequal sample sizes (one group sample size = 16 and other group = 8) and unequal variances. This study demonstrated greater adjusted power for Welch t-test followed by nonparametric bootstrap t (pooled t-statistic) under normal samples while Welch t-test followed by nonparametric bootstrap t (pooled t-statistic and Welch t-statistic) under non-normal samples. Similar to our study, this study also revealed greater unadjusted power with nonparametric bootstrap t (pooled t-statistic) in any situations as compared with others. In our study, we found bootstrap F-test performed similar or slightly better than the F-test in all conditions. These tests provided greater statistical power as compared with Kruskal–Wallis test. For comparing more than two means, authors [39] have evaluated the performance of F-test compared with bootstrap F-test under equal and unequal sample sizes with unequal variances. They found that bootstrap F-test provides similar or favorable results compared with the standard F-test. Furthermore, authors have recommended the bootstrap procedures as compared with standard nonparametric tests for comparing two or more means. However, this study used an unpooled error resampling method for determining empirical distribution under the null hypothesis as opposed to our study. Furthermore, this study showed only a few simulations under very limited conditions (unequal sample sizes and unequal variances).

A study compared type I error and power across unpaired *t*-test, Welch *t*-test, and rank *t*-test for extremely small sample size studies (<7) under various conditions using extensive simulations [12]. The results of this study were found very consistent to our study. For example, for the sample size of six (three per group), the power was 0.46 and 0.37 for *t*-test and Welch *t*-test, respectively, under normal distributions. In our study, the power was 0.46 and 0.36 for *t*-test and Welch *t*-test, respectively, under normal distributions. This study concluded that unpaired *t*-test should be preferred over Welch *t*-test and rank *t*-test in all conditions except for unequal variances and unequal sample sizes. We also found unpaired *t*-test provides better power performance than Welch t-test and Wilcoxon rank sum test, except for unequal variances and unequal sample size in small and extremely small sample size studies. In contrast to Winter [12] and findings from our simulation studies, a study demonstrated a greater power for Welch *t*-test compared with regular *t*-test for unequal variances [20].

Konietschke and Pauly compared the performance of nonparametric bootstrap paired *t*-test and permutation test with paired *t*-test [23]. They have extensively studied the type I error probabilities of nonparametric bootstrap paired *t*-test for sample sizes 7, 10, and 20. However, the power performance was only assessed for studies with a sample of 10 or 20 pairs with normal and lognormal distributions. In this study, authors showed improved power with nonparametric bootstrap paired *t*-test using counterintuitive resampling method compared with regular paired *t*-test under non-normality. Similar to this study, we also observed marked improvement in power for nonparametric bootstrap paired *t*-test compared with paired *t*-test under normal and skew normal distributions for large correlation (>70%) studies. Moreover, this study also demonstrated that the permutation test with counterintuitive resampling method outperforms nonparametric bootstrap paired *t*-test for lognormal distribution. We also found that the exact permutation *t*-test for comparison of two independent means yields higher power compared with our proposed nonparametric bootstrap *t*-test for lognormal data with sample size 20. Our study found remarkably higher power with Wilcoxon rank sum test compared with other alternatives for extreme variable lognormal data and Cauchy distributed data. However, this study did not compare the performance of Wilcoxon signed-rank test with permutation *t*-test or bootstrap paired *t*-test under this condition. Smuker *et al.* [46] provided comparison of randomization test, Wilcoxon signed-rank test, sign test, bootstrap *t*-test, and paired *t*-test on information retrieval evaluation data to compare difference between paired means. They found that randomization, bootstrap, and *t*-tests provided similar ability to detect significant difference between means. This study suggested discontinuing the use of Wilcoxon signed-rank test and sign test. The sample size in this study was 18,820 pairs, and this study did not evaluate performance of any tests in different conditions. Further, this study used mean difference as a test statistic and computed *p*-value using unpooled resampling method.

In the first data example for comparing percent seizures between two groups, we observed similar and significant *p*-values using all tests except from the exact Wilcoxon rank sum test. This is expected as the SDs were very similar between two groups. For a paired scenario, paired *t*-test, permutation *t*-test, and nonparametric bootstrap paired *t*-test provided similar *p*-values for mild variable data, while nonparametric bootstrap paired *t*-test and Wilcoxon signed-rank test provided similar *p*-values for large variable data. This confirms that nonparametric bootstrap paired *t*-test provides reliable results in any conditions. The *p*-values obtained using nonparametric bootstrap *t*-test and Welch *t*-test were found to be very similar, however, different than the unpaired *t*-test, permutation *t*-test, and exact Wilcoxon rank sum test for unequal group sizes and unequal variances in the second data example. This confirms that the nonparametric bootstrap *t*-test is useful for data with unequal sample sizes and large variance. Our simulations also demonstrated that exact standard nonparametric test has no power at all for extremely small sample size studies. Further, we did not find any power advantages in any situations with exact nonparametric test or permutation test compared with our proposed bootstrap test except for heavy-tailed distributions. We recommend using exact Wilcoxon test for Cauchy and extreme variance lognormal distributed data as opposed to permutation test. For less variable lognormal data with sample size 20 or higher, an exact permutation *t*-test may be preferred.

The selection of different test statistic in bootstrap tests may be explored for small sample size studies. We only explored comparisons of three means using nonparametric bootstrap *F*-test. However, our R code can be easily extended for many means. The performance of bootstrap test was only assessed for normal and skew normal distributions for comparison of paired and more than two independent means. In such situations, the benefits of using bootstrap test for other non-normal distributions may also be explored. Further, for paired situations, we have assessed performance of tests in large correlated data only as the proposed test referred as counterintuitive method has already been explored for small correlation studies [23]. The proposed bootstrap test, which is more suitable for comparing equality of distributions, may be compared with the bootstrap test that directly compares the means between groups.

In summary, our extensive simulations and analysis of real data examples showed that unpaired *t*-test is robust for small sample sizes studies and skew normal data if the sample sizes in two groups are similar. For unequal sample sizes, unequal variances, and skew normal data, Welch *t*-test is superior to Wilcoxon rank sum test followed by *t*-test, and permutation *t*-test but inferior to nonparametric bootstrap *t*-test. Mostly, permutation *t*-test produces considerable power advantages over exact Wilcoxon rank sum test for normal or skew normal distributed data with equal sample sizes but inferior to nonparametric bootstrap *t*-test and Student's *t*-test. Unpaired *t*-test should not be preferred for non-normal distributions. The proposed pooled non-parametric bootstrap *t*-test should be preferred for comparing two means, or for validating the findings of unpaired *t*-test for small sample size studies, especially for unequal sample sizes, unequal variances, and non-normal distributions. In general, data from lognormal distribution should be dealt carefully and analyzed with multiple tests (exact Wilcoxon rank sum *t*-test, nonparametric bootstrap *t*-test, and exact permutation test). For Cauchy distribution, Wilcoxon rank sum test should be preferred. For high correlated paired data, nonparametric bootstrap paired *t*-test is superior to paired *t*-test. Paired *t*-test and nonparametric bootstrap paired *t*-test are appropriate for less variable data. For extreme variable data, nonparametric bootstrap paired *t*-test and exact Wilcoxon signed-rank test produce similar results. We strongly suggest using the proposed nonparametric bootstrap test as an alternative to standard parametric, nonparametric, and permutation tests in small sample size studies because (1) it provides reliable and more powerful results in many conditions, (2) it does not require any assumptions as opposed to other alternatives, (3) it is feasible and easy to use with standard statistical software, and (4) it avoids confusion of selecting tests between parametric and nonparametric procedures. On a final note, this study recommends the use of proposed bootstrap test in small sample size experimental or randomized studies, but does not recommend conducting small size studies.

Appendix A:

A.1. Nonparametric bootstrap t-test

```
data < −read.table(“C:/data.csv”,header = T , sep = “,”)        # read the data in as data
val < − data$Y                                                  # quantitative dependent variable
grp < − data$X                                                  # binary group variable
bootR =1000                                                     # Number of bootstrap replicates
bootstrap = function (val,grp,bootR){
n1 = sum(grp==1)
n2 = sum(grp==2)
t.values = numeric(bootR)                                       #Store the test statistics values

for (j in 1:bootR) {
group1 = sample(val, size = n1, replace = T)                    #Sampling with replacement for group
                                                                1
group2 = sample(val, size = n2, replace = T)                    #Sampling with replacement for group
                                                                2
    if (sd(group1)==0 & sd(group2)==0) {t.values[j] = NA}       #If Standard Deviations are 0 assign NA
else {t.values[j] = t.test(group1,group2,var.equal = T, na.rm = T) #Saving the test statistics
$statistic}
}
p.boot = mean(abs(t.values) > =abs(t.test(val ~ grp)$statistic) , na.rm = T) #Obtaining the p-value
p.boot
}
bootstrap(val,grp,bootR) # function will produce the p-value
```

A.2. Nonparametric bootstrap paired t-test

```
data < −read.table(“C:/data.csv”,header = T , sep = “,”) # Read the data in as
data
sg1 = data$Y1
sg2 = data$Y2
val < − c(sg1,sg2)                                          # Quantitative dependent variable
n = length(data$Y1)
bootR =1000                                                 # Number of bootstrap replicates
bootstrap = function (val,bootR){
t.values = numeric(bootR)                                   #Store the test statistics values
for (j in 1:bootR) {
        group1 = sample(val, size = n, replace = T) #Sampling with replacement
for group 1
        group2 = sample(val, size = n, replace = T) #Sampling with replacement
for group 2
if (sd(group1)==0 & sd(group2)==0) {t.values[j] = NA}      #If Standard Deviations are 0
                                                            assign NA
else {t.values[j] = t.test(group1,group2,paired = TRUE)$statistic}#Saving the
test statistics
}
p.boot = mean(abs(t.values) > =abs(t.test(sg1,sg2,paired = TRUE)$statistic),  #Obtaining the p-value
na.rm = T)
p.boot
}
bootstrap(val,bootR)
```

A.3. Nonparametric bootstrap F-test for comparison of three means

```
data < −read.table(“C:/data.csv”,header = T , sep = “,”) #read the data in as data
val < − data$Y                                                  # Quantitative dependent
                                                                variable
grp < − data$X                                                  # Polytomous group variable
bootR =1000                                                     # Number of bootstrap
                                                                replicates
bootstrap = function (val,grp,bootR){
n1 = sum(grp==1)
n2 = sum(grp==2)
n3 = sum(grp==3)
f.values = numeric(bootR)                                       #Store the test statistics
                                                                values
for (j in 1:bootR) {
        group1 = sample(val, size = n1, replace = T) #Sampling with replacement for
group 1
        group2 = sample(val, size = n2, replace = T) #Sampling with replacement for
group 2
        group3 = sample(val, size = n3, replace = T) #Sampling with replacement for
group 3
if (sd(group1)==0 & sd(group2)==0 & sd(group3)==0) {f.values[j] = NA}   #If Standard Deviations are 0
                                                                assign NA
else {
p.boot = mean(abs(f.values) > =abs(anova(lm(val ~ grp))$”F value” [1]),na.
rm = TRUE) #Obtaining the p-value
p.boot
}
bootstrap(val,grp,bootR)
```

Acknowledgements

We would like to thank Patrick Tarwater for useful discussions about the bootstrap mechanism. We also would like to thank the Editor and three anonymous reviewers for their careful review and helpful suggestions. The second example included in the study reported here is a result of secondary analyses of data from clinical trial conducted as part of the National Drug Abuse Treatment Clinical Trials Network (CTN) sponsored by the National Institute on Drug Abuse (NIDA). Specifically, data from CTN-0005 (MI (Motivational Interviewing) to Improve Treatment Engagement and Outcome in Subjects Seeking Treatment for Substance Abuse) was included. CTN databases and information are available at http://datashare.nida.nih.gov.

References

1. Siegal S. Nonparametric statistics. *The American Statistician* 1957; **11**:13–19.
2. Janusonis S. Comparing two small samples with an unstable, treatment-independent baseline. *Journal of Neuroscience Methods* 2009; **179**:173–178.
3. Altman DG, Gore SM, Gardner MJ, Pocock SJ. Statistical guidelines for contributors to medical journals. *British Medical Journal (Clinical Research Ed.)* 1983; **286**:1489–1493.
4. Siegal S, Castellan NJ. Nonparametric Statistics for the Behavioral Sciences. McGraw-Hill: New York, 1988.
5. Mundry R, Fischer J. Use of statistical programs for nonparametric tests of small samples often leads to incorrect P values: examples from animal behavior. *Animal Behavior* 1998; **56**:256–259.
6. Barber JA, Thompson SG. Analysis of cost data in randomized trials: an application of the non-parametric bootstrap. *Statistics in Medicine* 2000; **19**:3219–3236.
7. Hall P, Martin M. On the bootstrap and two sample problems. *Australian Journal of Statistics* 1988; **30A**:179–192.
8. Jekel JF, Katz DL, Elmore JG. Epidemilogy, Biostatistics, and Preventive Medicine. W.B. Saunders Company: Philidelphia, 2001.
9. Weber M, Sawilowsky SS. Comparative power of the independent *t*, permutation *t*, and Wilcoxon tests. *Journal of Modern Applied Statistical Methods* 2009; **8**:10–15.
10. Bridge PD, Sawilowsky SS. Increasing physicians' awareness of the impact of statistics on research outcomes: comparative power of the *t*-test and and Wilcoxon rank-sum test in small samples applied research. *Journal of Clinical Epidemiology* 1999; **52**:229–235.
11. Tanizaki H. Power comparison of non-parametric tests: small-sample properties from Monte Carlo experiments. *Journal of Applied Statistics* 1997; **24**:603–632.
12. Winter JCF. Using the Student's *t*-test with extremely small sample sizes. *Practical Assessment, Research & Evaluation* 2013; **18**.
13. Sawilowsky SS, Hillman SB. Power of the independent samples *t* test under a prevalent psychometric measure distribution. *Journal of Consulting and Clinical Psychology* 1993; **60**:240–243.
14. Zimmerman DW, Zumbo BD. Parametric alternatives to the student *t* test under violation of normality and homogeneity of variance. *Perceptual and Motor Skills* 1992; **74**:835–844.
15. Wampold BE, Drew CJ. Theory and Application of Statistics. Mcgraw-Hill College: New York, 1990.
16. Tomkins CC. *An introduction to non-parametric statistics for health scientists University of Alberta Health Sciences Journal* 2006:3.
17. Lumley T, Diehr P, Emerson S, Chen L. The importance of the normality assumption in large public health data sets. *Annual Review of Public Health* 2002; **23**:151–169.
18. Zhou XH, Melfi CA, Hui SL. Methods for comparison of cost data. *Annals of Internal Medicine* 1997; **127**:752–756.
19. Sokal RR, Rohlf FJ. Introduction to Biostatistics. W.H. Freeman: New York, 1987.
20. Lansing L. Bootstrapping versus the Student's *t*: The Problems of Type I Error and Power: Lehigh University; 1999.
21. Kulkarni S. A comparison of type I error rates for the bootstrap contrast with the *t* test and the roburst rank order test for various sample sizes and variances. *Theses and Dissertations* 1993.
22. Kang Y, Harring JR. Investigating the Impact of Non-normality, Effect Size, and Sample Size on Two-Group Comparison Procedures: An Emperial Study. University of Maryland: College Park.
23. Konietschke F, Pauly M. Bootstrapping and permuting paired *t*-test type statistics. *Statistics and Computing* 2014; **24**:283–296.
24. Janssen A, Pauls T. A Monte Carlo comparison of studentized bootstrap and permutation tests for heterscedastic two sample problems. *Computational Statistics* 2005; **20**:369–383.
25. Ahad NA, Abdullah S, Lai CH, Ali NM. Relative power performance of *t*-test and bootstrap procedure for two sample. *Pertanika Journal of Science & Technology* 2012; **20**:43–52.
26. Conover WJ, Iman RL. Rank transformations as a bridge between parametric and non-parametric statistics. *The American Statistician* 1981; **35**:124–129.
27. Fisher RA. The design of experiments. *Edinburgh* 1935.
28. Berry KJ, Johnston JE. Mielke PW. Permutation methods: Computational Statistics, 2011.
29. Good PI. Permutation, Parametric, and Bootstrap Tests of Hypothesis (3rd edn): New York, 2005.
30. Md E. Permutation methods: a basis for exact inference. *Statistical Science* 2004; **19**:676–685.
31. Berger VW. Pros and cons of permutation tests in clinical trials. *Statistics in Medicine* 2000; **19**:1319–1328.
32. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. Chapman and Hall, 1993.
33. Phipson B, Smyth GK. Permutation *p*-values should never be zero: calculating exact *p*-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology* 2010; **9**:1–12.

34. Hesterberg T, Moore DS, Monaghan S, Clipson A. Epstein R. Bootstrap Methods and Permutation Tests: New York, 2005.
35. Pauly M, Brunner E, Konietschke F. Asymptotic permutation tests in general factorial designs. *Journal of the Royal Statistical Society Series B* 2015; **77**:461–473.
36. Corcoran CD, Mehta CR. Exact level and power of permutation, bootstrap, and asymptotic tests of trend. *Journal of Modern Applied Statistical Methods* 2002; **2**:42–51.
37. Ludbrook J, Dudley H. Why permutation tests are superior to t and F tests in biomedical research. *The American Statistician* 1998; **52**:127–133.
38. Lu M, Chase G, Li S. Permutation tests and other tests statistics for ill behaved data. *Communications in Statistics-Theory and Methods* 2001; **30**:1481–1496.
39. Boos DD, Brownie C. Bootstrap *p*-values for tests of nonparametric hypotheses. *Institute of Statistics Mimeo Series No* 1919; **1988**.
40. Lunneborg CE. Bootstrap Applications for the Behavioral Sciences. University of Washington: C.E. Lunneborg, 1987.
41. Azzalini A, Dalla VA. The multivariate skew-normal distribution. *Biometrika* 1996; **83**:715–726.
42. Azzalini A, Capitanio A. The Skew-Normal and Related Families. Cambridge University Press: IMS Monograph series, 2014.
43. Carroll KM, Ball SA, Nich C, Martino S, Frankforter TL, Farentinos C, Kunkel LE, Mikulich-Gilbertson SK, Morgenstern J, Obert JL, Polcin D, Snead N, Woody GE, National Institute on Drug Abuse Clinical Trials Network. Motivational interviewing to improve treatment engagement and outcome in individuals seeking treatment for substance abuse: a multisite effectiveness study. *Drug and Alcohol Dependence* 2006; **81**:301–312.
44. Janssen A, Pauls T. How do bootstrap and permutation tests work? *Annals of Statistics* 2003; **768-806**.
45. Adams DC, Anthony CD. Using randomization techniques to analyze behavioral data. *Animal Behavior* 1996; **54**:733–738.
46. Smucker MD, Allan J, Carterette B. A comparison of statistical significance tests for information retrieval evaluation. *ACM Conference on information and knowledge management* 2007; **2007**:623–632.

Supporting information

Additional Supporting Information may be found online in the supporting information tab for this article.