Eur. J. Stat. 6 (2026) 8
doi: 10.28924/ada/stat.6.8

Parametric Versus Non-Parametric Statistics: A Case Study in Traditional and Alternative Medicine Research with the Development of SDA4AMR Web Application

Julara Chumnaul*, Tasneem Sarong, Atittaya Promruangchot

Division of Computational Science, Faculty of Science, Prince of Songkla University, Songkhla, Thailand

julara.c@psu.ac.th, 6310210535@psu.ac.th, 6310210649@email.psu.ac.th

* *Correspondence: julara.c@psu.ac.th*

Abstract. This study investigated the performance of parametric and non-parametric tests, specifically, the one-sample t-test and the Wilcoxon Signed-Rank (WSR) test, through simulation-based evaluations of type I error rates and statistical power across varying sample sizes and effect sizes. Complementing the simulation study, we conducted a systematic review of empirical articles published in two journals in traditional and alternative medicine to assess the usage of statistical methods and the reporting of assumption checks. Simulation results revealed that the t-test generally maintained acceptable type I error rates under Bradley's criterion, especially when sample sizes were 20 or greater. In contrast, the WSR test frequently exhibited inflated error rates, particularly with larger samples. In terms of power, the t-test consistently outperformed the WSR test, though both achieved satisfactory power (≥ 0.8) under appropriate conditions. These findings underscore the importance of context-specific test selection, striking a balance between robustness and statistical sensitivity. Furthermore, the journal review revealed that, although a majority of articles employed inferential statistics, assumption checking was rarely reported, even for commonly used methods such as t-tests, ANOVA, and chi-square tests. This lack of transparency raises concerns about the validity of statistical conclusions drawn in the literature. Therefore, to support better statistical practice, the Smart Data Analysis Web Application for Alternative Medicine Research (SDA4AMR) was developed. This tool facilitates the selection of appropriate tests, automatic assumption checking, and interpretable outputs. By enhancing accessibility and encouraging methodological rigor, SDA4AMR aims to improve research quality and reproducibility in the field of traditional and alternative medicine.

1. Introduction

In the data analysis stage, especially when applying statistical inferences (parameter estimates and hypothesis testing), researchers must ensure that the assumptions underlying their statistical tests are met in order to obtain accurate results. To do so, researchers need to have

Received: 2 Jan 2026.

Key words and phrases. user-friendly software; web application; data analysis; assumptions checking; alternative medicine research.

a solid understanding of various statistical concepts, especially the assumptions underlying the tests they plan to use, as well as the techniques for verifying these assumptions. For example, when comparing the averages of two separate groups with the independent samples t-test, it is crucial to first check for normal distribution and equal variances across groups [1] using the Shapiro-Wilk and F tests, respectively. Similarly, to compare averages between more than two groups with one-way ANOVA, it is necessary to evaluate each group for normal distribution and equal variances [2], which can be done using the Shapiro-Wilk and Bartlett tests. However, many researchers, especially those who are not statisticians, often lack knowledge about these prerequisites and how to test for them before proceeding with their data analysis. This issue is prevalent in many studies, showing that a minority of research explicitly examines the underlying assumptions and mentions assumption checks related to the statistics used in their publications [3–8]. If the statistical method is used where assumptions are not met, it can lead to various issues, such as unreliable test statistics and their associated p-values [9], statistical errors, and biased estimates, with impacts ranging from minor to severe [1,10–20]. Olsen (2003) and Choi (2005) [21,22] underscore the importance of adhering to assumptions and caution that disregarding them can seriously compromise data analysis, potentially leading to false and unreliable conclusions. Moreover, researchers without statistical knowledge also struggle to choose a suitable statistical method for their data when the underlying assumptions are not satisfied.

In statistics, data analysis is divided into parametric and nonparametric methods. Parametric methods rely on certain assumptions about the underlying distribution of the data, typically assuming that the data follows the normal distribution. These methods are powerful and efficient when the assumptions are met, providing precise estimates and predictions. In contrast, nonparametric methods do not assume any specific distribution for the data. They are more flexible and can be applied when the assumptions needed for parametric methods are violated. Therefore, nonparametric methods are particularly useful when dealing with small sample sizes, ordinal data, or any data that does not meet the normality assumption.

In medical research, it is often noted that the collected data does not follow a normal distribution [23–25], and the sample sizes are often small. This deviation from normality can arise from various factors, such as the nature of the data, cultural practices, and specific characteristics of the populations being studied. Consequently, researchers frequently face challenges when using parametric statistical methods, which depend on the assumption of normally distributed data. Given these challenges, nonparametric methods offer a practical alternative for data analysis in medical research. Nonparametric methods do not require the assumption of normality, making them suitable for data that is skewed, contains outliers, or is ordinal. These methods offer flexibility and robustness, which makes them ideal for various research scenarios where parametric methods may not be appropriate. However, despite the advantages of nonparametric methods, many researchers may lack the necessary knowledge and skills to test for underlying assumptions and select suitable statistical techniques for their data. This gap in expertise can lead to the

misuse of statistical methods, potentially compromising the validity and reliability of research findings.

Nonparametric and parametric tests in biomedical research are important topics that have been widely discussed among researchers due to their significant impact on research outcomes. For example, Vickers (2005) [26] examined parametric versus non-parametric statistics in analyzing randomized trials with non-normally distributed data. The findings indicated that ANCOVA is the preferred method for analyzing randomized trials with both baseline and post-treatment measurements. However, in some extreme cases, ANCOVA may be less powerful than the Mann-Whitney test. Then, Kitchen (2009) [27] examined both nonparametric and parametric tests for the location and found that nonparametric tests, such as the Wilcoxon rank sum test (WRST) and Mann-Whitney U-test (MWUT) are effective alternatives to parametric tests when dealing with skewed, asymmetrical, multimodal, or heavy-tailed data, especially in small samples. Moreover, when data are normally distributed, and all of the other assumptions are met, using WRST or MWUT results in relatively little loss in power, and there can be substantial gains in reliability when the assumptions for parametric tests are violated. Therefore, nonparametric tests are a reliable choice for primary analysis. The comparison of nonparametric and parametric tests in biomedical research was also studied by Stojanović et al. (2018) [28], who found that nonparametric procedures are beneficial in many situations, and necessary in individuals, especially when the assumptions of a parametric test are not met. In such cases, a nonparametric test can be used as an alternative.

This study has four main objectives. The first objective is to compare the performance of commonly used parametric and nonparametric tests when the data does not follow a normal distribution. The second objective is to investigate the verification of the underlying assumptions of statistical methods utilized in medical research, specifically focusing on a case study in Thai traditional and alternative medicine research. Finally, the fourth objective is to develop a web application called Smart Data Analysis for Alternative Medicine Research (SDA4AMR) to assist researchers with limited statistical knowledge in analyzing data within the field of alternative medicine research.

2. Methodology

This study employed a two-pronged methodological approach to evaluate the performance and practical application of statistical testing methods. Section 2.1 presents a simulation study designed to assess the effectiveness of parametric and non-parametric tests under various data conditions. Section 2.1.2 involves a systematic review of academic journal articles to examine how assumptions underlying statistical methods are reported and addressed in empirical research.

2.1. **Experiment I: Simulation study.**

This experiment aims to investigate the visual analog scale (VAS) commonly used in Thai traditional medicine research. It is often observed that the

VAS does not follow a normal distribution. Typically, the VAS measures 100 mm in length, and when measured with a precision of 1 mm, it results in a scale with 101 points $(0, 1, ..., 100)$. In our study, we will report the VAS on a scale from 0 to 10 in centimeters.

According to Heller et al. (2016) [29], the distribution relevant to Visual Analog Scales (VAS) is the beta distribution. This distribution is bounded within a finite interval of (0, 1) and can be either symmetric or skewed. Therefore, we generate VAS using the beta distribution in two distinct patterns: left-skewed and right-skewed (see Figure 1). Subsequently, we conducted hypothesis testing to compare the mean VAS score with a predetermined value and evaluated the performance of two statistical testing methods: the t-test and the Wilcoxon signed-rank (WSR) test, representing parametric and non-parametric approaches, respectively. The objective was to assess whether the t-test remains effective when its underlying assumptions are violated, and to determine which of the two methods performs better under varying conditions.

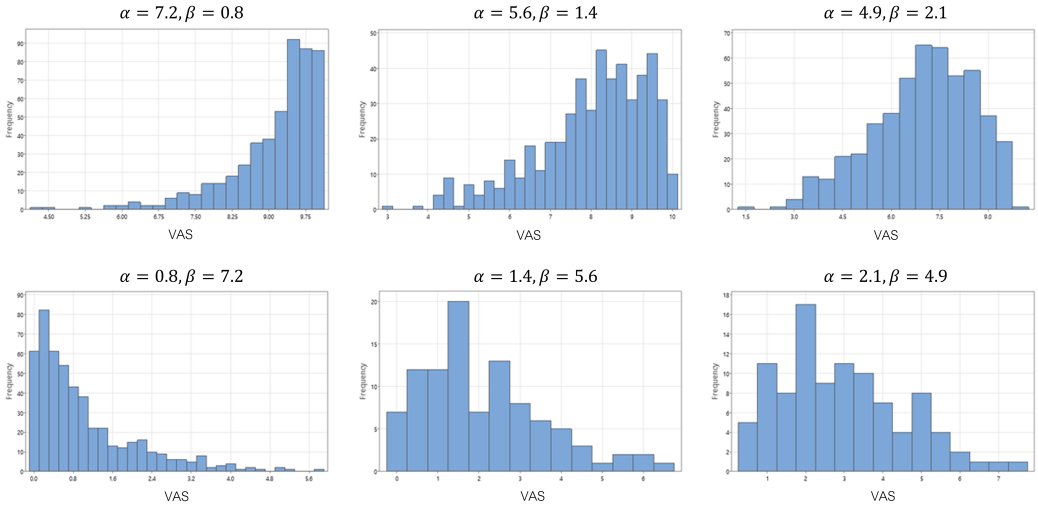

FIGURE 1. Generate data based on the beta distribution to represent VAS.

2.1.1. *Scope of the simulation study.* For this experiment, we carefully defined the scope of the simulation study to ensure a comprehensive evaluation of the statistical methods under investigation. The following parameters and conditions were established to guide the simulation design and analysis:

1. The data used in this study follow a beta distribution with parameter values selected to represent both left-skewed and right-skewed distributions.

2. The sample sizes (n) considered in the analysis are 10, 15, 20, 25, and 30.

3. The hypothesis tests conducted are two-tailed tests with a significance level of $\alpha = 0.05$.

4. The performance of the tests is evaluated based on their ability to control the probability of type I error and their statistical power.

5. Each scenario was simulated 10,000 times to ensure the reliability and robustness of the results.

2.1.2. *Simulation study procedure.* To evaluate the performance of the t-test and the Wilcoxon signed-rank test under various data conditions, a simulation-based approach was employed. The simulation procedure involved generating data under both the null and alternative hypotheses, applying the statistical tests, and assessing their performance based on type I error rates and statistical power. The detailed steps are as follows:

Step 1: Define the hypothesis of the test as follows:

$$H_0 : \mu = \mu_0 \text{ versus } H_1 : \mu = \mu_1.$$

Step 2: Generate datasets according to the scenarios defined in the scope of the study. To evaluate the probability of type I error, data were generated under the null hypothesis (H_0), while power was assessed using data generated under the alternative hypothesis (H_1).

Step 3: Calculate the test statistics for both the t-test and the Wilcoxon signed-rank test for each dataset.

Step 4: Compare the calculated test statistics with the corresponding critical values obtained using statistical software. Based on this comparison, determine whether to reject or fail to reject the null hypothesis (H_0).

Step 5: Repeat Steps 1 to 3 a total of 10,000 times for each scenario to ensure the robustness of the results.

Step 6: Record the number of times the null hypothesis was rejected in each scenario.

Step 7: Calculate the following for both tests: the empirical type I error rate ($\hat{\alpha}$), which is the proportion of rejections under H_0, and the empirical power ($1 - \hat{\beta}$), which is the proportion of rejections under H_1.

Step 8: Compare the empirical type I error rates with Bradley's criterion [30] at a significance level of 0.05. If the empirical type I error rate falls within the interval $[0.025, 0.075]$, the test is considered to have acceptable control over type I error.

Step 9: Compare the empirical power of the two tests. If both tests adequately control the type I error rate, the test with higher power is considered more efficient.

2.2. **Experiment II: Review of research articles that examine and report assumptions of statistical methods.** For the second experiment, we conducted a review of research articles that examined and reported the underlying assumptions of commonly used parametric statistical methods in medical research, with a specific focus on studies related to the Thai traditional and alternative medicine. The research methodology for this experiment is organized into three main subsections: Sample, Coding, and Data Analysis. Each subsection outlines the specific methods and processes employed to conduct the research effectively and accurately.

2.2.1. *Sample.* This study examined research articles in Thai traditional and alternative medicine, focusing on publications in the Thai-Journal Citation Index (TCI) database. We specifically selected two journals for inclusion: the Thai Traditional Medicine Research Journal and the Journal of Thai Traditional and Alternative Medicine. We synthesized 258 articles from

each issue published between 2018 and 2023 in these two journals to identify related research. We documented studies that employed at least one statistical method, including the independent t-test, paired t-test, one-way analysis of variance (ANOVA), repeated measures ANOVA, chi-squared test for association, and Pearson correlations.

2.2.2. *Coding.* The coding scheme used in this study aimed to evaluate how thoroughly statistical assumptions were reported in research articles. Since many studies employed multiple statistical tests, two distinct coding standards stringent and lenient were applied to ensure consistent evaluation [18]. The lenient standard was considered satisfied if any assumption related to the statistical methods used was mentioned in the study; this means that even if the study did not comprehensively address all assumptions of a particular statistical test, the mere mention of one assumption was sufficient to meet the lenient criterion.

In contrast, the stringent standard held reporting to a higher threshold. All relevant assumptions for the statistical methods used had to be explicitly stated and verified in the study; this could include assumptions such as normality, equal variances, independence of observations, linearity, or any other conditions necessary for the tests' validity.

The coding scheme, presented in Table 1, served as the framework for systematically evaluating how well studies adhered to these two standards. By applying both lenient and stringent criteria, the coding scheme provided a comprehensive assessment of the statistical rigor in reporting and checking assumptions across multiple studies.

2.2.3. *Keyword identification.* Keywords related to statistical assumptions were identified using the search function in Adobe Acrobat to assist with coding. The keywords included terms such as *assumption*, *normality*, *equal variances*, *sphericity*, *independence*, *outlier*, *linearity*, and *expected frequency*. This automated keyword search was the initial step in determining whether a study had reported the assumptions underlying its statistical tests.

2.2.4. *Manual review.* After conducting the initial keyword search, the articles were examined more thoroughly. We reviewed the Research Methodology and Results sections and any relevant endnotes or appendices to gather additional evidence regarding whether the studies explicitly considered and checked their statistical assumptions. This manual review ensured that, even if certain assumptions were not captured during the keyword search, they could still be identified through careful reading. The coding scheme presented in Table 1 served as a framework for systematically evaluating how well the studies adhered to these two standards. By applying both lenient and stringent criteria, the coding scheme offered a comprehensive assessment of the statistical rigor in reporting and checking assumptions across multiple studies.

3. Results

3.1. **Result of Experimental I.** This section presents the empirical evaluation of the parametric and non-parametric methods under various simulation settings. Table 2 reports the

TABLE 1. Coding scheme.

Part	No	Item/check	Answer
I	1	Article title	.....
	2	Journal	.....
	3	Year	.....
	4	Volume/Issue number	.....
	5	Is this paper empirical?	☐ Yes ☐ No
	6	Which type of empirical research was adopted?	☐ Non-empirical ☐ Empirical
	7	Which type(s) of statistics was/ were used?	☐ Descriptive ☐ Inferential ☐ Both
II	8	Whether or not each of the following is used as statistical procedure?	
		• Independent sample t-test	☐ Yes ☐ No
		• Paired t-test	☐ Yes ☐ No
		• One-way ANOVA	☐ Yes ☐ No
		• Chi-square test for association	☐ Yes ☐ No
		• Pearson correlation	☐ Yes ☐ No
		• Repeated measures ANOVA	☐ Yes ☐ No
		• Other statistical methods	☐ Yes ☐ No
III	9	For each of the statistical procedures used, is assumption-checking reported?	☐ Yes ☐ No
	10	Which assumptions were reported?	.....
	11	Does this paper report assumption-checking according to a stringent standard?	☐ Yes ☐ No
	12	Does this paper report assumption-checking according to a lenient standard?	☐ Yes ☐ No

empirical type I error rates across different scenarios, providing insight into the methods' ability to maintain nominal significance levels under the null hypothesis. In contrast, Table 3 through 8 summarize the empirical powers under a range of alternative hypotheses, highlighting the sensitivity and effectiveness of each method in detecting true effects.

According to Bradley's criterion, an acceptable empirical type I error rate should be within the interval $[0.025, 0.075]$, indicating adequate control of the test size. Across all settings, the t-test generally maintains type I error rates close to or slightly above the nominal level, as shown in Figure 2. For sample sizes $n \geq 20$, most t-test error rates approach the Bradley range, indicating improved control with increasing sample size. However, for smaller sample sizes ($n = 10$ or 15),

the t-test shows some inflation (e.g., 0.0915 for $\mu_0 = 9$), suggesting potential liberal behavior. In contrast, the WSR test consistently exhibits higher type I error rates than the t-test, especially as the sample size increases. For instance, at $n = 30$, WSR error rates exceed 0.08 in several cases and reach as high as 0.1427 for $\mu_0 = 1$, significantly outside the Bradley range. Even at lower sample sizes, the WSR test tends to be liberal, with values well above the 0.06 threshold.

TABLE 2. Empirical type I error rates for testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ at 5% significance level.

n	**Method**	$\mu_0 = 9$	$\mu_0 = 8$	$\mu_0 = 7$	$\mu_0 = 3$	$\mu_0 = 2$	$\mu_0 = 1$
10	t-test	0.0915	0.0672	0.0577	0.0565	0.0669	0.0899
	WSR test	0.0803	0.0623	0.0537	0.0529	0.0595	0.0795
15	t-test	0.0857	0.0594	0.0527	0.0574	0.0625	0.0840
	WSR test	0.1022	0.0641	0.0522	0.0542	0.0687	0.1024
20	t-test	0.0745	0.0603	0.0542	0.0533	0.0615	0.0707
	WSR test	0.1127	0.0697	0.0557	0.0546	0.0728	0.1105
25	t-test	0.0722	0.0578	0.0571	0.0525	0.0549	0.0694
	WSR test	0.1289	0.0718	0.0620	0.0582	0.0723	0.1235
30	t-test	0.0669	0.0579	0.0523	0.0531	0.0561	0.0702
	WSR test	0.1398	0.0817	0.0606	0.0617	0.0799	0.1427

Regarding Tables 3-8, the power of both tests increases with larger sample sizes and greater differences between the true mean and the hypothesized value in all scenarios, reflecting the stronger ability to detect a true effect. Generally, the t-test exhibits slightly higher power than the WSR test, particularly when sample sizes are small and the data follow normal distributions, making it more sensitive under these conditions. However, the WSR test still performs comparably and may be preferred when the normality assumption is questionable. Across all parameter settings, both tests reach high power (often above 0.8) when the sample size is at least 25 and the difference in means is sufficiently large (see Figure 3).

In addition, this study also presents an illustrative example of testing the difference in central tendency between two independent groups, in a case where the results from parametric and non-parametric tests yield different conclusions. The example utilizes VAS (Visual Analogue Scale) data from patients in Groups 1 and 2. The dataset and corresponding analysis results are presented in Tables 9-11 and Figure 4.

The analysis of VAS scores between the two groups demonstrates how data distribution can influence statistical test outcomes. Although the independent sample t-test yields a p-value of 0.053, which is slightly above the conventional 0.05 threshold and therefore leads to retaining the null hypothesis, the Mann-Whitney U test gives a p-value of 0.003, indicating a significant difference (see Table 11). This discrepancy suggests that the assumptions underlying the t-test (normality and equal variances) may not be met, as indicated by the skewed boxplot and unequal spreads in the histograms (see Figure 4).

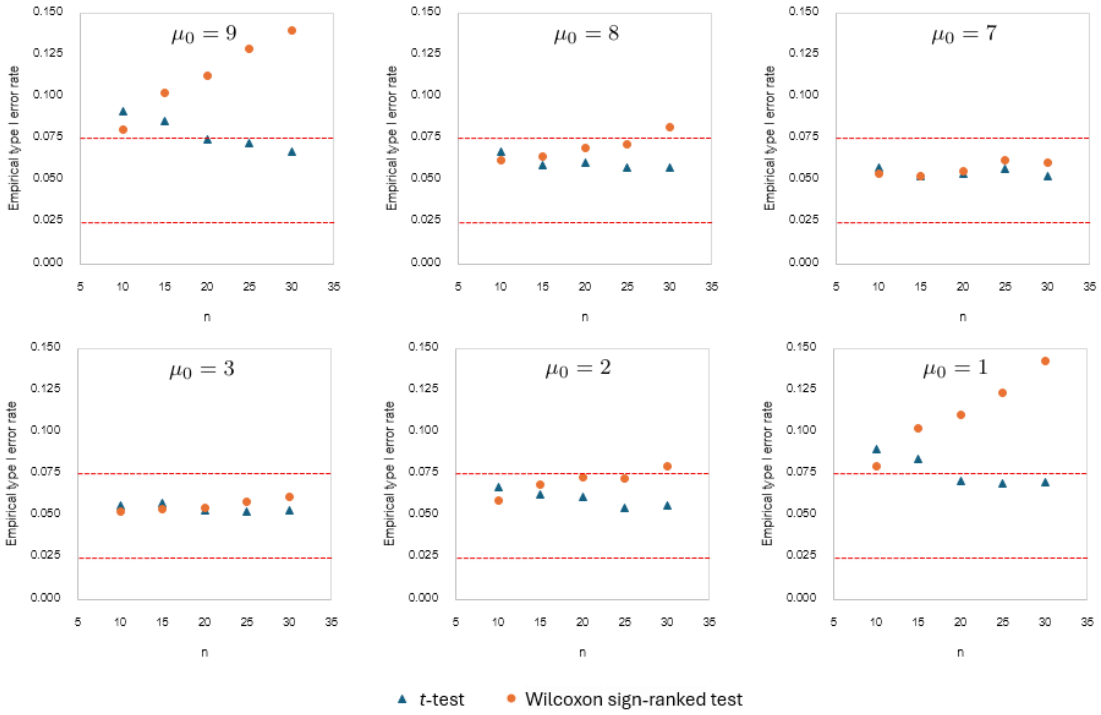

FIGURE 2. Empirical type I error rates comparisons between the t-test (blue triangle) and the Wilcoxon signed-rank test (orange circle) across varying hypotheses and sample sizes.

TABLE 3. Empirical powers for testing $H_0 : \mu = 9$ versus $H_1 : \mu = \mu_1$ at 5% significance level.

n	**Method**	$\mu_1 = 8$	$\mu_1 = 8.5$	$\mu_1 = 9$	$\mu_1 = 9.5$	$\mu_1 = 10$
10	t-test	0.7406	0.3908	0.0964	0.1870	0.9585
	WSR test	0.6950	0.3536	0.0844	0.1729	1.0000
15	t-test	0.8840	0.4891	0.0783	0.3990	0.9994
	WSR test	0.8877	0.5211	0.0981	0.2895	1.0000
20	t-test	0.9534	0.5774	0.0778	0.5778	1.0000
	WSR test	0.9599	0.6468	0.1158	0.3805	1.0000
25	t-test	0.9788	0.6546	0.0675	0.7335	1.0000
	WSR test	0.9833	0.7446	0.1271	0.4838	1.0000
30	t-test	0.9918	0.7239	0.0719	0.8339	1.0000
	WSR test	0.9967	0.8333	0.1438	0.5701	1.0000

t denotes one-sample t-test, and WSR denotes Wilcoxon sign-ranked test.

In practice, this highlights the importance of checking data assumptions before choosing a test. When the data are not normally distributed or variances are unequal, as is likely in this case, non-parametric tests like the Mann-Whitney U test are more appropriate. Researchers

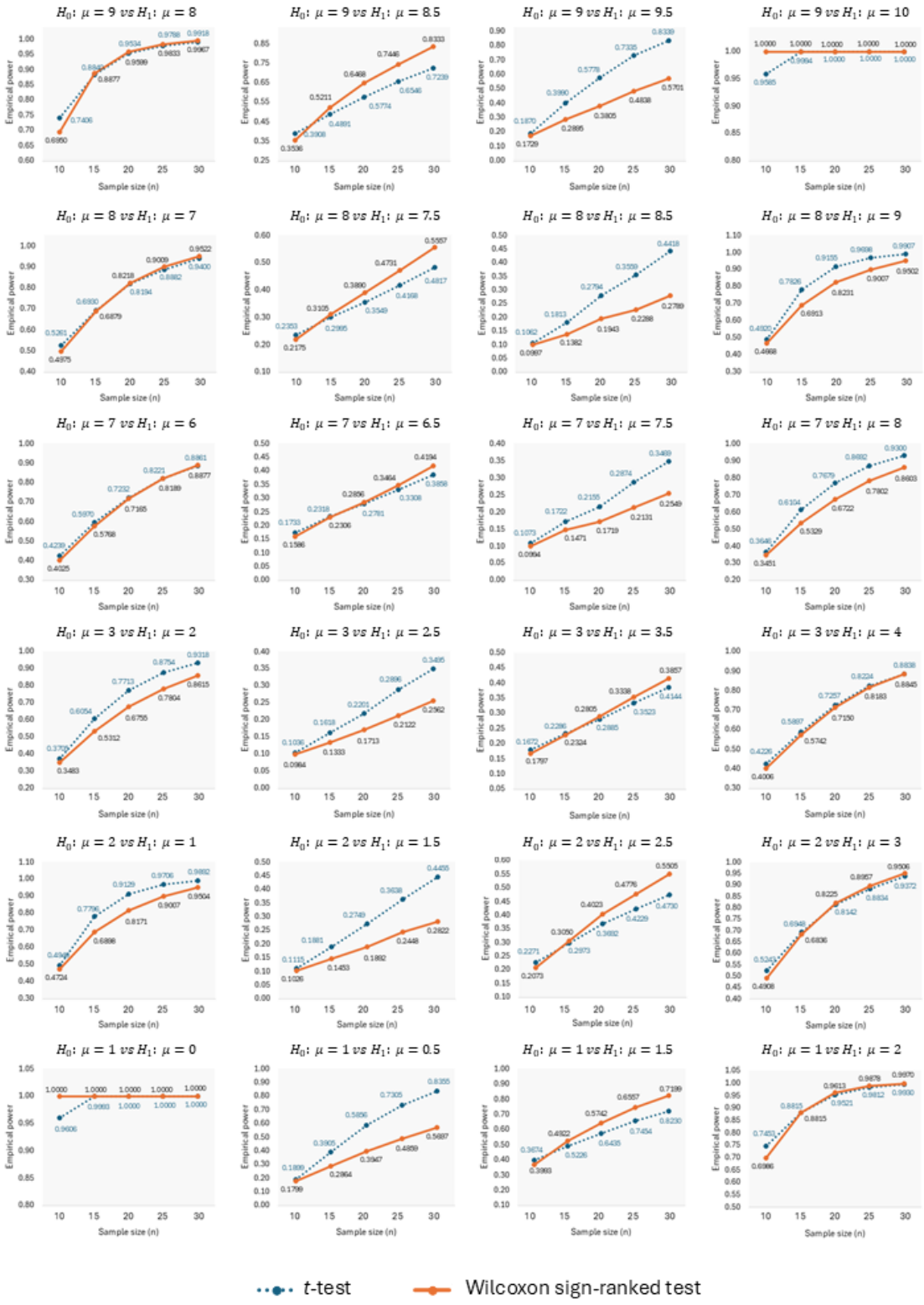

FIGURE 3. Empirical power comparisons between the t-test (blue dashed line) and the Wilcoxon signed-rank test (orange solid line) across varying hypotheses and sample sizes.

TABLE 4. Empirical powers for testing $H_0 : \mu = 8$ versus $H_1 : \mu = \mu_1$ at 5% significance level.

n	Method	$\mu_1 = 7$	$\mu_1 = 7.5$	$\mu_1 = 8$	$\mu_1 = 8.5$	$\mu_1 = 9$
10	t-test	0.5261	0.2353	0.0709	0.1062	0.4920
	WSR test	0.4975	0.2175	0.0631	0.0997	0.4668
15	t-test	0.6930	0.2995	0.0647	0.1813	0.7826
	WSR test	0.6879	0.3105	0.0685	0.1382	0.6913
20	t-test	0.8194	0.3549	0.0628	0.2794	0.9155
	WSR test	0.8218	0.3890	0.0749	0.1943	0.8231
25	t-test	0.8882	0.4168	0.0561	0.3559	0.9698
	WSR test	0.9009	0.4731	0.0753	0.2288	0.9007
30	t-test	0.9400	0.4817	0.0553	0.4418	0.9907
	WSR test	0.9522	0.5557	0.0815	0.2789	0.9502

TABLE 5. Empirical powers for testing $H_0 : \mu = 7$ versus $H_1 : \mu = \mu_1$ at 5% significance level.

n	Method	$\mu_1 = 6$	$\mu_1 = 6.5$	$\mu_1 = 7$	$\mu_1 = 7.5$	$\mu_1 = 8$
10	t-test	0.4239	0.1733	0.0550	0.1073	0.3646
	WSR test	0.4025	0.1586	0.0515	0.0994	0.3451
15	t-test	0.5970	0.2318	0.0598	0.1722	0.6104
	WSR test	0.5768	0.2306	0.0577	0.1471	0.5329
20	t-test	0.7232	0.2781	0.0561	0.2155	0.7679
	WSR test	0.7165	0.2856	0.0573	0.1719	0.6722
25	t-test	0.8221	0.3308	0.0584	0.2874	0.8692
	WSR test	0.8189	0.3464	0.0609	0.2131	0.7802
30	t-test	0.8861	0.3858	0.0561	0.3469	0.9300
	WSR test	0.8877	0.4194	0.0610	0.2549	0.8603

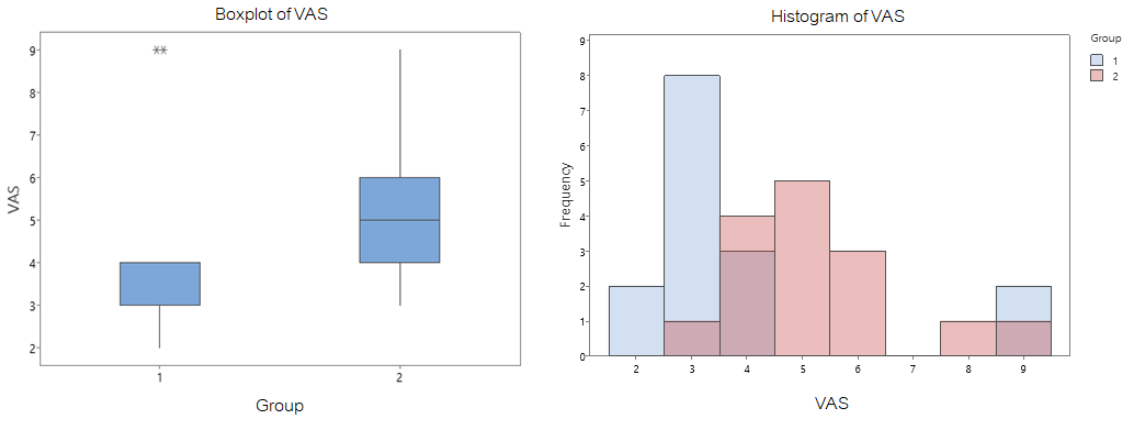

FIGURE 4. Boxplot and histogram of VAS scores by group.

TABLE 6. Empirical powers for testing $H_0 : \mu = 3$ versus $H_1 : \mu = \mu_1$ at 5% significance level.

n	Method	$\mu_1 = 2$	$\mu_1 = 2.5$	$\mu_1 = 3$	$\mu_1 = 3.5$	$\mu_1 = 4$
10	*t*-test	0.3705	0.1036	0.0534	0.1797	0.4226
	WSR test	0.3483	0.0984	0.0497	0.1672	0.4006
15	*t*-test	0.6054	0.1618	0.0574	0.2324	0.5897
	WSR test	0.5312	0.1333	0.0536	0.2286	0.5742
20	*t*-test	0.7713	0.2201	0.0538	0.2805	0.7257
	WSR test	0.6755	0.1713	0.0544	0.2885	0.7150
25	*t*-test	0.8754	0.2896	0.0526	0.3338	0.8224
	WSR test	0.7804	0.2122	0.0580	0.3523	0.8183
30	*t*-test	0.9318	0.3495	0.0542	0.3857	0.8838
	WSR test	0.8615	0.2562	0.0611	0.4144	0.8845

TABLE 7. Empirical powers for testing $H_0 : \mu = 2$ versus $H_1 : \mu = \mu_1$ at 5% significance level.

n	Method	$\mu_1 = 1$	$\mu_1 = 1.5$	$\mu_1 = 2$	$\mu_1 = 2.5$	$\mu_1 = 3$
10	*t*-test	0.4940	0.1115	0.0695	0.2271	0.5243
	WSR test	0.4724	0.1026	0.0615	0.2073	0.4908
15	*t*-test	0.7796	0.1881	0.0648	0.2973	0.6948
	WSR test	0.6898	0.1453	0.0689	0.3050	0.6836
20	*t*-test	0.9129	0.2749	0.0576	0.3692	0.8142
	WSR test	0.8171	0.1892	0.0696	0.4023	0.8225
25	*t*-test	0.9706	0.3638	0.0569	0.4229	0.8834
	WSR test	0.9007	0.2448	0.0738	0.4776	0.8957
30	*t*-test	0.9892	0.4455	0.0542	0.4730	0.9372
	WSR test	0.9504	0.2822	0.0784	0.5505	0.9506

should routinely examine the data visually (e.g., using boxplots or histograms) and statistically before selecting a test.

3.2. **Result of Experimental II.** According to Table 12, out of 258 articles, 233 (90.31%) are empirical, representing a significant majority. In contrast, 25 articles (9.69%) are non-empirical. Table 13 also highlights the dominance of empirical research in the studied journals, with a proportion exceeding 0.8 ($z = 4.14$, $p < 0.001$); this indicates that most articles in the *Thai Traditional Medicine Research Journal* and the *Journal of Thai Traditional and Alternative Medicine* use statistical analysis to support their research findings.

For the empirical articles, Table 14 reveals that 158 (67.81%) employed inferential statistics in their analysis, while 75 articles (32.19%) did not. Furthermore, Table 14 also shows that the proportion of empirical articles using inferential statistics is significantly greater than 50%

TABLE 8. Empirical powers for testing $H_0 : \mu = 1$ versus $H_1 : \mu = \mu_1$ at 5% significance level.

n	Method	$\mu_1 = 0$	$\mu_1 = 0.5$	$\mu_1 = 1$	$\mu_1 = 1.5$	$\mu_1 = 2$
10	t-test	0.9606	0.1899	0.0888	0.3993	0.7453
	WSR test	1.0000	0.1799	0.0776	0.3674	0.6986
15	t-test	0.9993	0.3905	0.0816	0.4922	0.8815
	WSR test	1.0000	0.2864	0.1011	0.5226	0.8815
20	t-test	1.0000	0.5856	0.0782	0.5742	0.9521
	WSR test	1.0000	0.3947	0.1128	0.6435	0.9613
25	t-test	1.0000	0.7305	0.0711	0.6557	0.9812
	WSR test	1.0000	0.4859	0.1257	0.7454	0.9878
30	t-test	1.0000	0.8355	0.0698	0.7199	0.9930
	WSR test	1.0000	0.5697	0.1412	0.8230	0.9970

TABLE 9. VAS scores of patients in Group 1 and 2.

Group	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	4	3	3	4	3	9	3	3	9	3	3	2	4	3	2
2	3	4	5	5	4	4	6	9	5	6	4	5	6	8	5

TABLE 10. Descriptive statistics of VAS by group.

Group	Mean	Std. Deviation	Median	IQR	Minimum	Maximum
1	3.87	2.167	3.00	1	2	9
2	5.27	1.580	5.00	2	3	9

TABLE 11. Analysis results from parametric and non-parametric tests at 5% significance level.

Method	Hypothesis	Statistic	p-value	Decision
t-test (Equal variances)	$H_0 : \mu_1 - \mu_2 = 0$	-2.022	0.053	Cannot reject H_0
	$H_1 : \mu_1 - \mu_2 \neq 0$			
Mann-Whitney U test	$H_0 : M_1 - M_2 = 0$	183.000	0.003*	Reject H_0
	$H_1 : M_1 - M_2 \neq 0$			

*Significant at the 5% level.

at the 0.05 significance level ($z = 5.44$, $p < 0.001$); this suggests that inferential statistics is a common approach in empirical research, likely because it allows researchers to make population-level inferences from samples, test hypotheses, and assess the reliability of their results. When examining the overall rate of assumption checking, only 2.33% of the studies explicitly mentioned verifying assumptions before conducting statistical analysis under the stringent standard (see

Table 15). However, this reporting rate slightly increased to 3.49% when using the lenient standard.

As shown in 16, the paired t-test emerges as the most frequently employed statistical method in articles published by the Thai Traditional Medicine Research Journal and the Journal of Thai Traditional and Alternative Medicine, accounting for 28.68% (74 articles) of the total. Other commonly used techniques include the independent sample t-test (15.89%, 41 articles), one-way ANOVA (12.02%, 31 articles), chi-square test (10.47%, 27 articles), repeated measures ANOVA (6.20%, 16 articles), and Pearson correlation (3.49%, 9 articles). For assumption checking, Table 16 reveals a generally low rate of reporting across these six statistical methods. Among the 74 studies using the paired t-test, only 6 (8.10%) reported checking assumptions. For the independent t-test, 6 out of 41 articles (14.62%) included such reporting. Similarly, assumption checks were observed in 6 out of 31 one-way ANOVA articles (19.35%), 1 out of 27 chi-square test articles (3.70%), 2 out of 16 repeated measures ANOVA articles (12.50%), and 2 out of 9 Pearson correlation studies (22.22%).

TABLE 12. Frequency and percentage of articles by the type of statistical analysis used.

Article	**Frequency**	**Percentage**
Empirical research	233	90.31
Employing inferential statistics	158	67.81
Without employing inferential statistics	75	32.19
Non-empirical research	25	9.69

TABLE 13. Results of testing the proportion of empirical research articles under the hypothesis $H_0 : p \leq 0.8$ against $H_1 : p > 0.8$.

Article	**Frequency** (%)	z	p**-value**
Empirical research	233 (90.31)	4.14	$< 0.001^*$
Non-empirical research	25 (9.69)		

*Significant at the 5% level.

TABLE 14. Results of testing the proportion of empirical research articles that use inferential statistics to analyze data under the hypothesis $H_0 : p \leq 0.5$ against $H_1 : p > 0.5$.

Empirical research	**Frequency** (%)	z	p**-value**
Employing inferential statistics	158 (67.81)	5.44	$< 0.001^*$
Without employing inferential statistics	75 (32.19)		

*Significant at the 5% level.

TABLE 15. Reporting practices in assumption-checking.

Type of assumption-checking	k	**Percentage**
Stringent standard	6	2.33
Lenient standard	9	3.49

*k is the number of studies.

TABLE 16. Assumption-checking across the focal statistics.

Statistical method	k_{total}	$k_{checked}$ **(Percentage)**
Paired t-test	74	6 (8.10)
Independent t-test	41	6 (14.62)
One-way ANOVA	31	6 (19.35)
Chi-square test	27	1 (3.70)
Repeated measures ANOVA	16	2 (12.50)
Pearson correlation	9	2 (22.22)
Other statistical methods	82	1 (1.22)

*k is the number of studies, and some studies employ multiple statistics.

A closer examination of the specific assumptions checked, as detailed in Table 17, reveals varied patterns. For the paired t-test, all six articles that performed assumption checks focused on the assumption of normality. Among the independent t-test studies, 3 (7.30%) assessed normality and 3 (7.31%) tested for equal variances, while none addressed independent errors. Within the one-way ANOVA group, only 1 article (3.22%) tested normality, whereas 5 (16.13%) verified equal variances, and none evaluated error independence. The chi-square test received minimal attention to assumptions, with only one study (3.70%) checking the minimum expected frequencies, and none addressing the independence of variables. For repeated measures ANOVA, two studies (12.50%) examined normality, but none tested for sphericity. Lastly, in the context of Pearson correlation, two studies (22.22%) checked for normality, while none assessed outliers or linearity. In summary, Table 17 highlights that checking for normality is the most commonly reported assumption across the tests where it is applicable. However, other assumptions such as equal variances, independent errors, minimum expected frequencies, sphericity, outliers, and linearity are less frequently reported, with several not checked at all in the studies reviewed. The overall low percentages suggest that assumption checking is not consistently performed or reported across the studies using these statistical methods.

4. DEVELOPMENT OF THE SMART DATA ANALYSIS WEB APPLICATION FOR ALTERNATIVE MEDICINE RESEARCH (SDA4AMR)

Based on the data analysis results of the previous section, the Smart Data Analysis Web Application for Alternative Medicine Research (SDA4AMR), accessible at https://jularatchumnaul.

TABLE 17. Individual assumptions reported for each statistical test.

Assumptions	k	**Percentage**
Paired t-test (k = 74)		
Normality	6	8.10
Independent t-test (k = 41)		
Normality	3	7.30
Equal variances	3	7.30
Independent errors	0	0.00
One-way ANOVA (k = 31)		
Normality	1	3.22
Equal variances	5	16.13
Independent errors	0	0.00
Chi-square (k = 27)		
Minimum expected frequencies	1	3.70
Independent variables	0	0.00
Repeated measures ANOVA (k = 16)		
Normality	2	12.50
Sphericity	0	0.00
Pearson correlation (k = 9)		
Normality	2	22.22
Outlier	0	0.00
Linearity	0	0.00

shinyapps.io/SDA4AMR/ (see Figure 5) was developed using the Shiny package of R software in the hope that it will be a tool to help researchers in the field of traditional and alternative medicine research in the data analysis step.

SDAAMR expands on the Smart Data Analysis V2 (SDA-V2) web application developed by Chumnaul and Sepherifar (2024) [31]. It includes additional statistical packages commonly used in traditional and alternative medicine research (RM ANOVA and simple logistic regression). The tool also offers a user manual in English, providing instructions on how to use each package and interpret and summarize analysis results (see Figure 6). Moreover, the functions and menu bars, as well as the web application display, are also in English, making them accessible to non-Thai researchers.

Like SDA-V2, the initial step in utilizing SDA4AMR involves preparing a dataset in csv format with UTF-8 encoding. Users should then upload this data table to the application. Following this, in the Variable(s) and parameters panel, users must input essential research information, such as the variables under study, the type of research hypothesis (two-sided or one-sided), the hypothesized value, and the desired significance level.

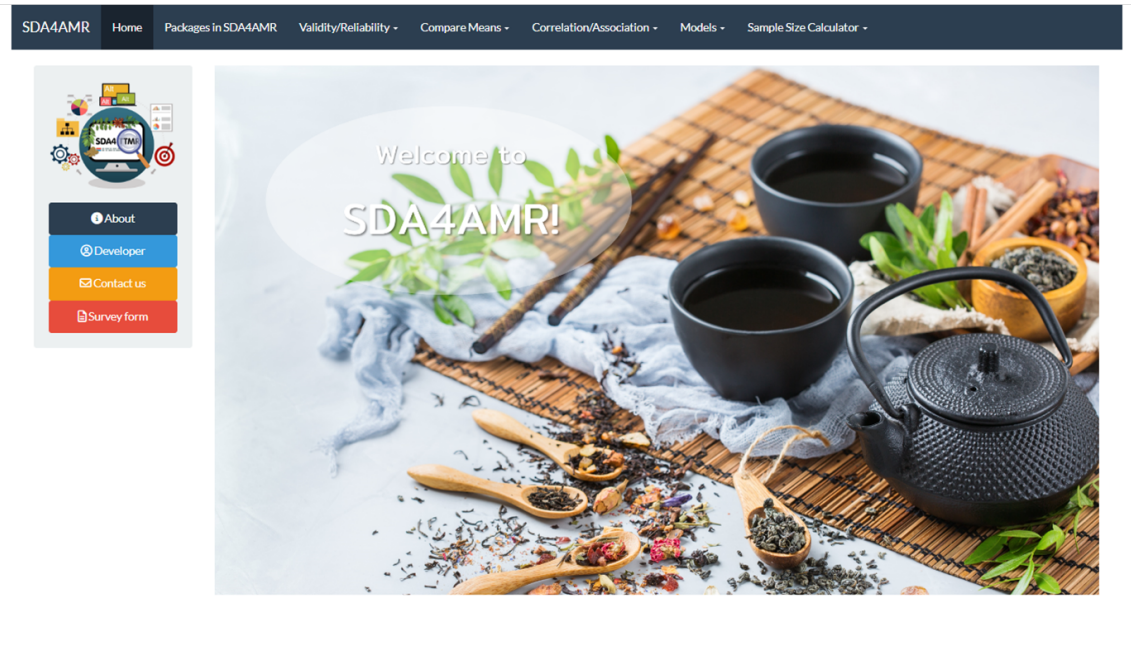

FIGURE 5. Smart Data Analysis Web Application for Alternative Medicine Research.

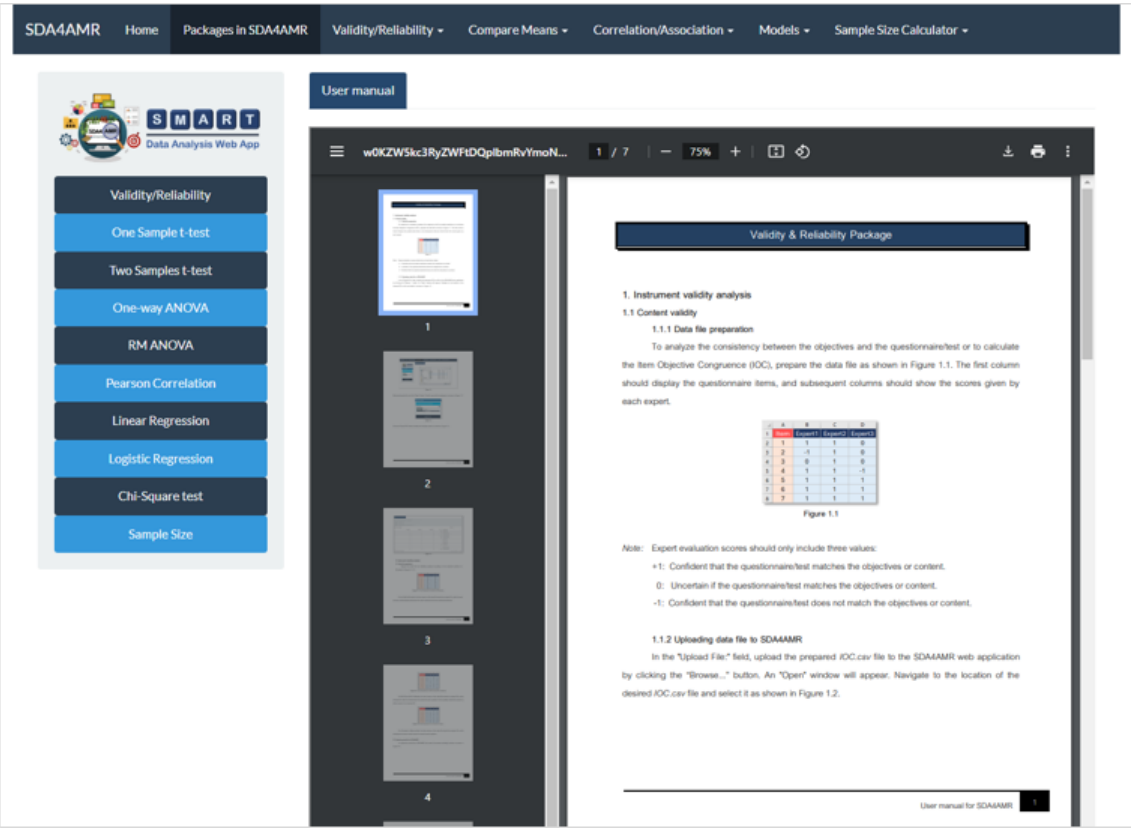

FIGURE 6. Statistical packages developed in SDA4AMR.

Once the data is in SDA4AMR, it is in good hands. The system automatically processes the data and associated details on an RStudio-hosted server. It starts by exploring and visualizing the data, then moves on to evaluating the underlying assumptions relevant to the parametric test and choosing a suitable statistical method.

5. Example demonstrating the SDA4AMR effectiveness

This section presents two case studies to illustrate the implementation and effectiveness of SDA4AMR. The first case study compares the stress reduction scores between two groups of volunteers undergoing different stress treatments. The second case study examines the impact of Triphala on the reduction of triglyceride levels.

5.1. **Example I: Stress treatments.** In this example, a researcher aims to evaluate the effectiveness of two stress reduction treatments. Twelve volunteers were assigned to an experimental group and received a manual-guided massage treatment for stress relief, while ten volunteers in the control group received standard care without specific stress reduction instructions. At the end of the experiment, participants rated their level of stress reduction on a scale from 0 to 10, with 0 indicating no reduction and 10 indicating maximum reduction. The stress reduction scores for both groups are displayed in Table 18.

Table 18. Volunteers' stress reduction scores.

Experimental group	**Control group**
8	2
8	3
7	5
8	6
9	4
8	5
9	4
7	6
9	4
8	5
7	
8	

In this illustrative example, we apply the *Two Samples t-test* package to determine whether a manual-guided massage treatment for stress relief reduces stress more than standard care without specific stress reduction instructions. Initially, users should prepare a data table of the stress reduction scores in csv format, as shown in Figure 7. Subsequently, this data table should be uploaded to SDA4AMR. Upon successful upload, users are required to identify the two groups for analysis by selecting *First group:* and *Second group:* and setting the comparison type to *Independent* under the option *Two populations are*. In line with the study's objective,

	A	B
1	Experimental_group	Control_group
2	8	2
3	8	3
4	7	5
5	8	6
6	9	4
7	8	5
8	9	4
9	7	6
10	9	4
11	8	5
12	7	
13	8	

FIGURE 7. Data table preparation for stress reduction scores.

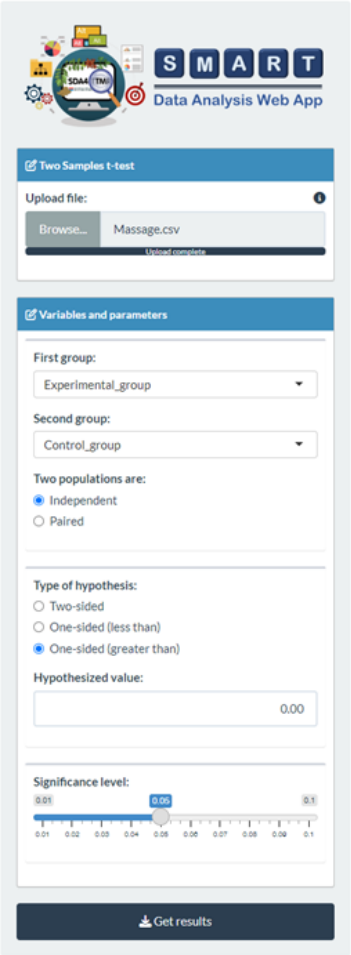

FIGURE 8. Data table uploaded and relevant research details provided.

users can hypothesize that a manual-guided massage treatment for stress relief is more effective than standard care. Therefore, the hypothesis type should be set to *One-sided (greater than)*

with a *Hypothesized value:* of *0.00*, and the significance level is set at 0.05. Figure 8 displays the data entry results.

Upon clicking the *Get results* button, SDA4AMR instantly displays all relevant analysis results across its various tabs. Figure 9 includes boxplots and histograms that illustrate the stress reduction scores for volunteers in each group. Meanwhile, Figure 10 provides basic statistics (sample size, mean, median, maximum, minimum, standard deviation, variance, and interquartile range) for the selected variables.

Results of assumptions checking for the related parametric test (two independent samples t-test) are shown in Figure 11. They indicate that the stress reduction scores for the experimental group are not normally distributed ($p = 0.020$), while those for the control group are normally distributed ($p = 0.445$). Consequently, the Wilcoxon rank sum test is automatically chosen and performed for the given data [32, 33] (see Figure 12). The hypothesis test result based on the Wilcoxon rank sum test, as presented in Figure 13, shows that stress reduction scores in the experimental group were significantly higher than those in the control group at the 5% significance level ($p = 0.000$), indicating that the manual-guided massage treatment effectively reduced stress.

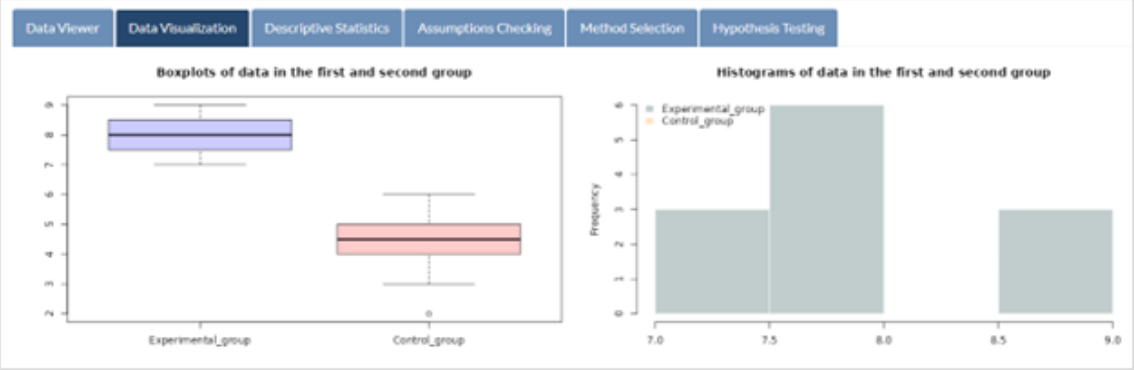

FIGURE 9. Boxplots and histograms of stress reduction scores.

5.2. **Example 2: Effectiveness of triphala recipe treatment.** Researchers conducted a study to explore the impact of Triphala on reducing triglyceride levels in 20 overweight or obese volunteers. Initially, they measured the participants' baseline triglyceride levels. Afterward, the volunteers were given 1800 milligrams of Triphala daily for 8 and 16 weeks. Triglyceride measurements were retaken at the 8-week and 16-week marks. The first 20 rows of data from this study are presented in Table 19.

In this example, the *RM ANOVA* (Repeated Measures ANOVA) package is used to evaluate the effect of Triphala treatment on triglyceride levels. Users begin by preparing a data table of triglyceride levels in csv format, and uploading it to SDA4AMR. After uploading, users specify the *Subject ID*, *Time point (or condition)*, and *Response variable* for analysis and set the significance level at 0.01.

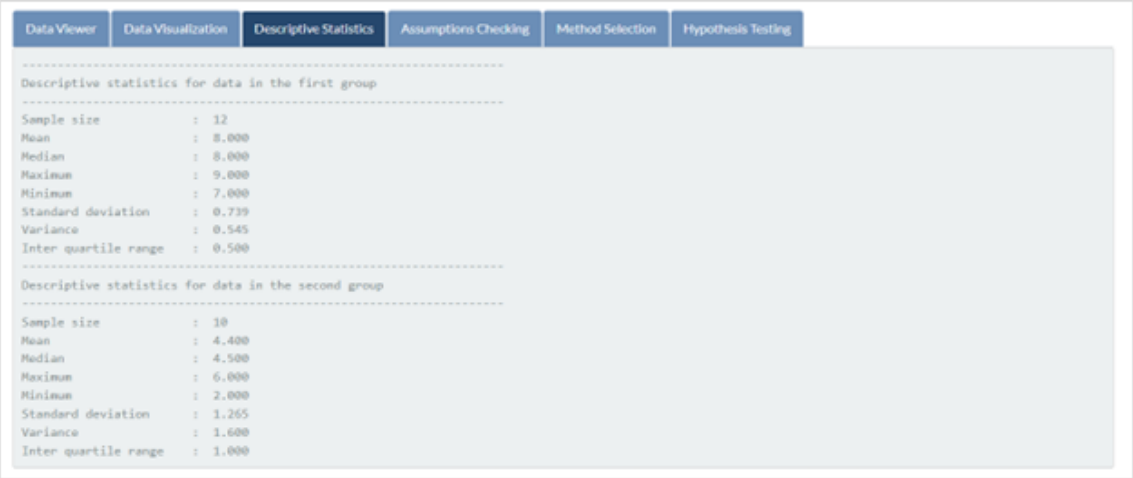

FIGURE 10. Basic statistics of stress reduction scores.

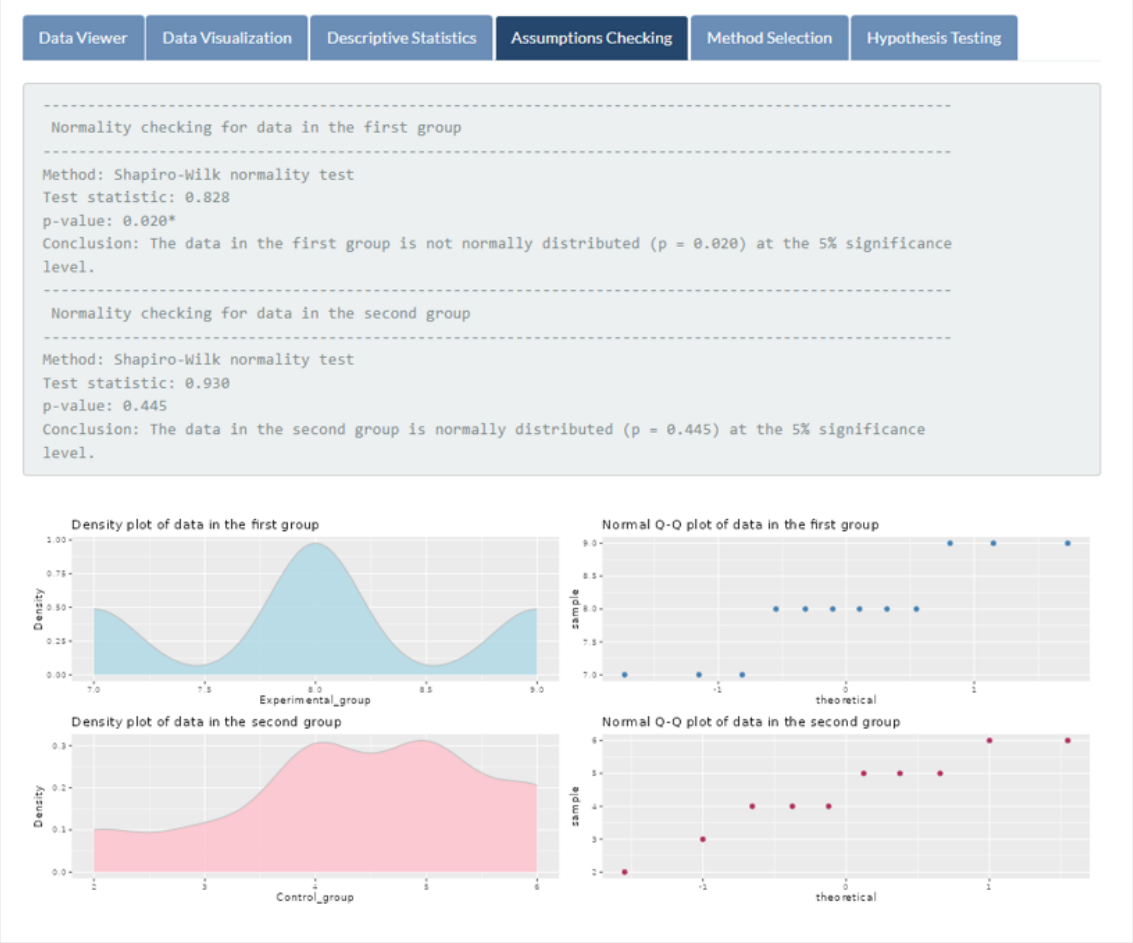

FIGURE 11. Assumptions checking for stress reduction scores.

FIGURE 12. Steps in choosing the appropriate statistical method for stress reduction scores.

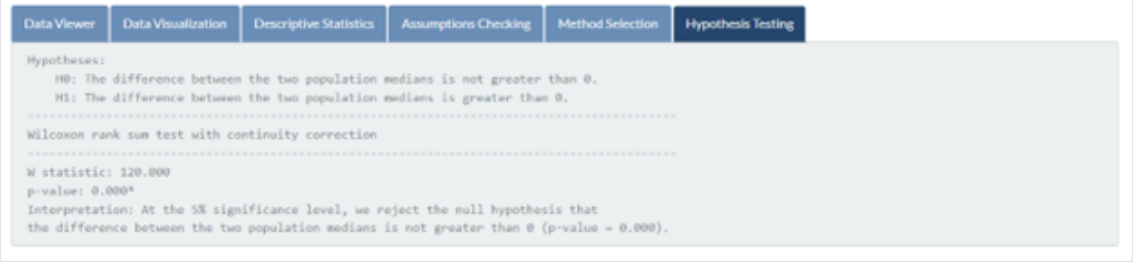

FIGURE 13. Results of the appropriate statistical method for stress reduction scores.

After clicking the *Get results* button, all relevant analysis results are immediately displayed on the SDA4AMR tabs, as presented in Figures 14-18. Figure 14 showcases multiple boxplots depicting triglyceride levels before treatment, after 8 weeks, and after 16 weeks, while Figure 15 displays the basic statistics of the response variable at each time point.

The assumption checks, shown in Figure 16, indicate that triglyceride levels at each time point are normally distributed at the 1% significance level. However, the data does not meet the sphericity assumption (p-value < 0.001). Consequently, Repeated Measures ANOVA with Greenhouse-Geisser correction is automatically applied (see Figure 17) [34–37]. The results of this analysis (Figure 18) show insufficient evidence to conclude a significant difference in mean triglyceride levels across time points at the 5% significance level (p-value $= 0.214$).

As the examples above illustrate, SDA4AMR offers a user-friendly and efficient solution for researchers without a formal background in statistics. Users are only required to upload their data and specify the relevant research details, after which SDA4AMR automates the processes of data exploration and visualization. The system evaluates the underlying assumptions associated

TABLE 19. First 20 observations from the simulated triglyceride levels study dataset.

ID	Gender	Weight	Height	BMI	Time	Triglyceride
1	Male	56.1	153	23.965	Before treatment	107
2	Female	99.8	178	31.499	Before treatment	211
3	Female	71.4	173	23.856	Before treatment	160
4	Female	72.9	161	28.124	Before treatment	149
5	Female	118	175	38.531	Before treatment	88
6	Female	58	156	23.833	Before treatment	86
7	Female	66.5	170	23.01	Before treatment	130
8	Female	65.2	160	25.469	Before treatment	134
9	Female	83.4	161	32.175	Before treatment	160
10	Male	59	154	24.878	Before treatment	71
11	Male	61.5	161	23.726	Before treatment	255
12	Male	94.4	166	34.258	Before treatment	118
13	Male	59.7	156	24.532	Before treatment	161
14	Male	80	178	25.249	Before treatment	135
15	Male	75	162	28.578	Before treatment	182
16	Male	93.5	170	32.353	Before treatment	65
17	Female	71.1	157	28.845	Before treatment	94
18	Female	84	168	29.762	Before treatment	155
19	Female	63.6	162	24.234	Before treatment	242
20	Female	55	152	23.805	Before treatment	303

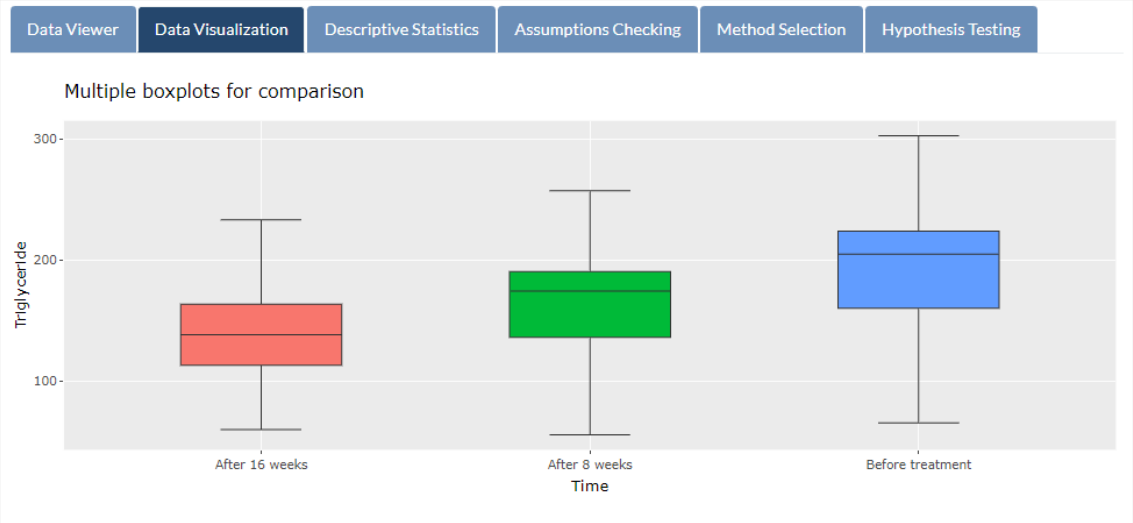

FIGURE 14. Multiple boxplots depicting triglyceride levels.

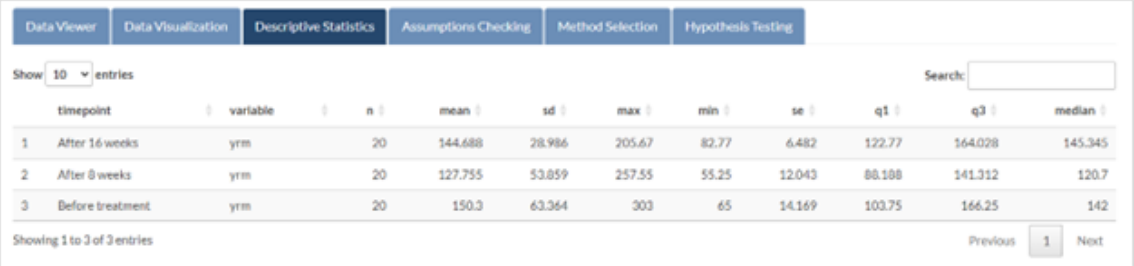

FIGURE 15. Basic statistics for triglyceride levels.

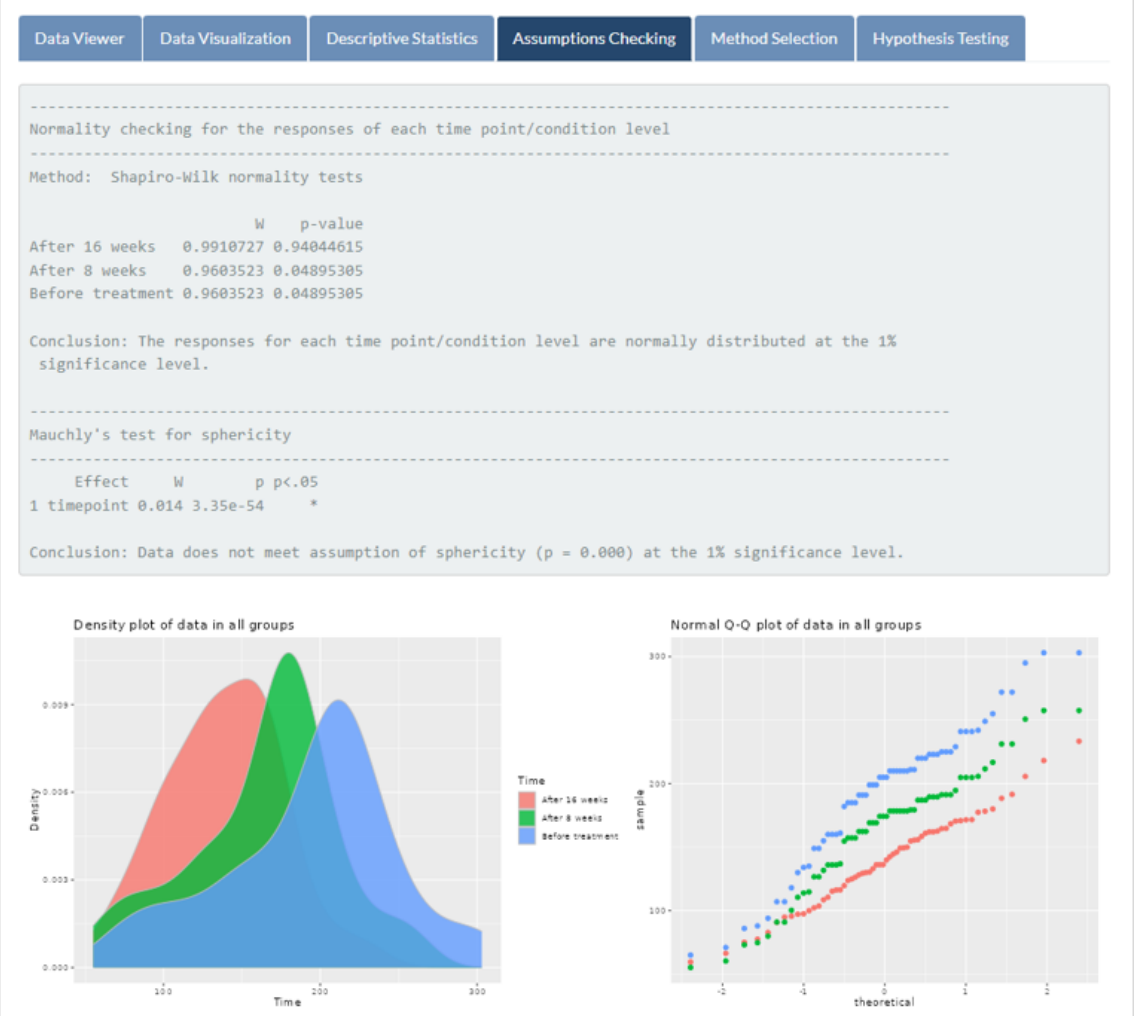

FIGURE 16. Results of normality and sphericity checking.

with parametric tests, identifies the most appropriate statistical methods for the given data, and presents the results in a clear and interpretable format. In contrast to conventional statistical software, which often requires users to manually assess assumptions and select suitable analytical techniques, SDA4AMR streamlines the workflow into a more accessible and integrated

FIGURE 17. Steps in choosing the appropriate statistical method for triglyceride levels data.

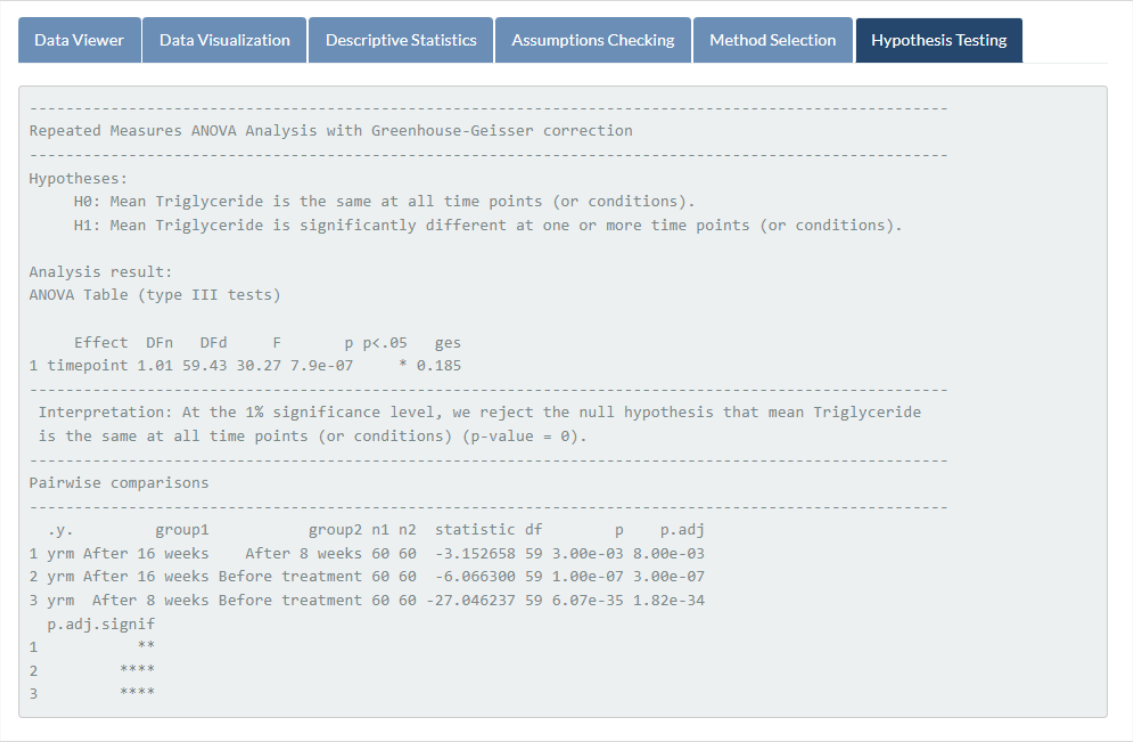

FIGURE 18. Results of the appropriate statistical method for triglyceride levels data.

process. For example, even a basic task such as comparing two means typically involves multiple procedural steps and a solid understanding of statistical principles.

6. Discussion and conclusion

This study examined the performance of parametric and non-parametric statistical tests by using the one-sample t-test and Wilcoxon Signed-Rank (WSR) test as representative methods of their respective categories. Through controlled simulations, we evaluated their empirical Type I error rates and statistical power under various sample sizes and effect sizes. Additionally, we assessed statistical practices in published empirical studies in traditional and alternative medicine research, focusing on the application and reporting of statistical methods, as well as assumption checking.

The findings from Experimental I highlight the importance of evaluating statistical test performance, particularly in terms of type I error control and statistical power. Based on Bradley's criterion, the t-test generally maintains acceptable type I error rates, especially as sample size increases. When $n \geq 20$, the t-test demonstrates error rates that are consistently within or near the acceptable interval [0.025,0.075]. In contrast, the Wilcoxon Signed-Rank (WSR) test shows a tendency to exceed this threshold, particularly for larger sample sizes, indicating liberal behavior and potential inflation of false positive rates. This finding suggests that while the WSR test may be a viable non-parametric alternative, caution is needed in its application, especially in hypothesis testing scenarios where type I error control is critical.

Power analysis further supports the t-test's efficacy in small to moderate sample sizes. It consistently outperforms the WSR test in detecting true differences in means, especially when the effect size is moderate to large. Nonetheless, the WSR test performs comparably in several settings and may be preferred in practice when data deviates from normality. Both tests exhibit satisfactory power levels (often exceeding 0.8) when sample sizes reach 25 or greater and the difference in means is sufficient. These findings underscore the importance of context-driven test selection, balancing robustness against assumption violations with the need for statistical sensitivity.

Experimental II sheds light on the current practices in statistical reporting within the Thai Traditional Medicine Research Journal and the Journal of Thai Traditional and Alternative Medicine. While the majority of articles (90.31%) are empirical and a significant proportion (67.81%) utilize inferential statistics, the rate of assumption checking remains alarmingly low. Even among frequently used methods, such as paired and independent t-tests, one-way ANOVA, and chi-square tests, verification of assumptions is seldom reported, with a rate often below 20% for any specific assumption. This lack of transparency can undermine the validity of statistical inferences, especially when researchers rely on parametric methods whose reliability hinges on assumptions such as normality, equal variances, and independence.

To address this gap, the Smart Data Analysis Web Application for Alternative Medicine Research (SDA4AMR) was developed as a practical tool to guide researchers through appropriate statistical practices. By incorporating assumption checks, model selection guidance, and clear output interpretation, the platform encourages a more rigorous approach to statistical analysis.

This tool not only builds on existing infrastructure (SDA-V2) but also enhances accessibility by providing English-language support and a user-friendly interface. Ultimately, the SDA4AMR application aims to promote reproducibility, improve research quality, and foster statistical literacy in the field of traditional and alternative medicine.

Competing interests: The authors declare that there is no conflict of interest regarding the publication of this paper.

References

[1] A.F. Hayes, L. Cai, Further Evaluating the Conditional Decision Rule for Comparing Two Independent Means, Br. J. Math. Stat. Psychol. 60 (2007), 217–244. https://doi.org/10.1348/000711005x62576.

[2] D.A. Kashy, M.B. Donnellan, R.A. Ackerman, D.W. Russell, Reporting and Interpreting Research in PSPB: Practices, Principles, and Pragmatics, Pers. Soc. Psychol. Bull. 35 (2009), 1131–1142. https://doi.org/10.1177/0146167208331253.

[3] R. Hoekstra, H. Kiers, A. Johnson, Are Assumptions of Well-Known Statistical Techniques Checked, and Why (Not)?, Front. Psychol. 3 (2012), 137. https://doi.org/10.3389/fpsyg.2012.00137.

[4] L. Plonsky, S. Gass, Quantitative Research Methods, Study Quality, and Outcomes: The Case of Interaction Research, Lang. Learn. 61 (2011), 325–366. https://doi.org/10.1111/j.1467-9922.2011.00640.x.

[5] L. Plonsky, Study Quality in Sla: An Assessment of Designs, Analyses, and Reporting Practices in Quantitative L2 Research, Stud. Second. Lang. Acquis. 35 (2013), 655–687. https://doi.org/10.1017/S0272263113000399.

[6] S. Lindstromberg, Inferential Statistics in Language Teaching Research: A Review and Ways Forward, Lang. Teach. Res. 20 (2016), 741–768. https://doi.org/10.1177/1362168816649979.

[7] Y. Hu, L. Plonsky, Statistical Assumptions in L2 Research: A Systematic Review, Second. Lang. Res. 37 (2019), 171–184. https://doi.org/10.1177/0267658319877433.

[8] S. Loewen, E. Lavolette, L.A. Spino, M. Papi, J. Schmidtke, et al., Statistical Literacy among Applied Linguists and Second Language Acquisition Researchers, TESOL Q. 48 (2013), 360–388. https://doi.org/10.1002/tesq.128.

[9] A.F. Ernst, C.J. Albers, Regression Assumptions in Clinical Psychology Research Practice—A Systematic Review of Common Misconceptions, PeerJ 5 (2017), e3323. https://doi.org/10.7717/peerj.3323.

[10] Y. Gel, W. Miao, J. Gastwirth, The Importance of Checking the Assumptions Underlying Statistical Analysis: Graphical Methods for Assessing Normality, Jurimetrics 46 (2005), 3–29.

[11] A.F. Hayes, L. Cai, Using Heteroskedasticity-Consistent Standard Error Estimators in OLS Regression: An Introduction and Software Implementation, Behav. Res. Methods 39 (2007), 709–722. https://doi.org/10.3758/bf03192961.

[12] A.F. Zuur, E.N. Ieno, C.S. Elphick, A Protocol for Data Exploration to Avoid Common Statistical Problems, Methods Ecol. Evol. 1 (2010), 3–14. https://doi.org/10.1111/j.2041-210x.2009.00001.x.

[13] P.J. Rosopa, M.M. Schaffer, A.N. Schroeder, Managing Heteroscedasticity in General Linear Models, Psychol. Methods 18 (2013), 335–351. https://doi.org/10.1037/a0032553.

[14] S. Troncoso Skidmore, B. Thompson, Bias and Precision of Some Classical ANOVA Effect Sizes When Assumptions Are Violated, Behav. Res. Methods 45 (2012), 536–546. https://doi.org/10.3758/s13428-012-0257-2.

[15] L.E. Barker, K.M. Shaw, Best (but Oft-Forgotten) Practices: Checking Assumptions Concerning Regression Residuals, Am. J. Clin. Nutr. 102 (2015), 533–539. https://doi.org/10.3945/ajcn.115.113498.

[16] A. Poncet, D.S. Courvoisier, C. Combescure, T.V. Perneger, Normality and Sample Size Do Not Matter for the Selection of an Appropriate Statistical Test for Two-Group Comparisons, Methodology 12 (2016), 61–71. https://doi.org/10.1027/1614-2241/a000110.

[17] A.F. Schmidt, C. Finan, Linear Regression and the Normality Assumption, J. Clin. Epidemiol. 98 (2018), 146–151. https://doi.org/10.1016/j.jclinepi.2017.12.006.

[18] Y. Hu, L. Plonsky, Statistical Assumptions in L2 Research: A Systematic Review, Second. Lang. Res. 37 (2019), 171–184. https://doi.org/10.1177/0267658319877433.

[19] U. Knief, W. Forstmeier, Violating the Normality Assumption May Be the Lesser of Two Evils, Behav. Res. Methods 53 (2021), 2576–2590. https://doi.org/10.3758/s13428-021-01587-5.

[20] G. Vallejo, M.P. Fernández, P. Rosário, Combination Rules for Homoscedastic and Heteroscedastic MANOVA Models from Multiply Imputed Datasets, Behav. Res. Methods 53 (2020), 669–685. https://doi.org/10.3758/s13428-020-01429-w.

[21] C.H. Olsen, Review of the Use of Statistics in Infection and Immunity, Infect. Immun. 71 (2003), 6689–6692. https://doi.org/10.1128/iai.71.12.6689-6692.2003.

[22] P.T. Choi, Statistics for the Reader: What to Ask Before Believ-Ing the Results, Can. J. Anesth. 52 (2005), R46. https://doi.org/10.1007/bf03023086.

[23] M. Hill, W.J. Dixon, Robustness in Real Life: A Study of Clinical Laboratory Data, Biometrics 38 (1982), 377–396. https://doi.org/10.2307/2530452.

[24] T. Micceri, The Unicorn, the Normal Curve, and Other Improbable Creatures., Psychol. Bull. 105 (1989), 156–166. https://doi.org/10.1037/0033-2909.105.1.156.

[25] P.D. Bridge, S.S. Sawilowsky, Increasing Physicians' Awareness of the Impact of Statistics on Research Outcomes: Comparative Power of the t-test and Wilcoxon Rank-Sum Test in Small Samples Applied Research, J. Clin. Epidemiol. 52 (1999), 229–235. https://doi.org/10.1016/s0895-4356(98)00168-1.

[26] A.J. Vickers, Parametric Versus Non-Parametric Statistics in the Analysis of Randomized Trials with Non-Normally Distributed Data, BMC Med. Res. Methodol. 5 (2005), 35. https://doi.org/10.1186/1471-2288-5-35.

[27] C.M. Kitchen, Nonparametric Vs Parametric Tests of Location in Biomedical Research, Am. J. Ophthalmol. 147 (2009), 571–572. https://doi.org/10.1016/j.ajo.2008.06.031.

[28] M. Stojanović, M. Andjelković-Apostolović, Z. Milošević, A. Ignjatović, Parametric versus Nonparametric Tests in Biomedical Research, Acta Medica Median. 57 (2018), 75–80. https://doi.org/10.5633/amm.2018.0212.

[29] G.Z. Heller, M. Manuguerra, R. Chow, How to Analyze the Visual Analogue Scale: Myths, Truths and Clinical Relevance, Scand. J. Pain 13 (2016), 67–75. https://doi.org/10.1016/j.sjpain.2016.06.012.

[30] J.V. Bradley, Robustness?, Br. J. Math. Stat. Psychol. 31 (1978), 144–152. https://doi.org/10.1111/j.2044-8317.1978.tb00581.x.

[31] J. Chumnaul, M. Sepehrifar, Smart Data Analysis V2: A User-Friendly Software for Non-Statisticians, PLOS ONE 19 (2024), e0297930. https://doi.org/10.1371/journal.pone.0297930.

[32] F. Wilcoxon, Individual Comparisons by Ranking Methods, Biom. Bull. 1 (1945), 80–83. https://doi.org/10.2307/3001968.

[33] W. Haynes, Wilcoxon Rank Sum Test, in: Encyclopedia of Systems Biology, Springer New York, 2013: pp. 2354–2355. https://doi.org/10.1007/978-1-4419-9863-7_1185.

[34] J. Zhao, C. Wang, S.C. Totton, J.N. Cullen, A.M. O'Connor, Reporting and Analysis of Repeated Measurements in Preclinical Animals Experiments, PLOS ONE 14 (2019), e0220879. https://doi.org/10.1371/journal.pone.0220879.

[35] J.W. Mauchly, Significance Test for Sphericity of a Normal n-Variate Distribution, Ann. Math. Stat. 11 (1940), 204–209. https://doi.org/10.1214/aoms/1177731915.

[36] S.W. Greenhouse, S. Geisser, On Methods in the Analysis of Profile Data, Psychometrika 24 (1959), 95–112. https://doi.org/10.1007/bf02289823.

[37] R. Rana, R. Singhal, V. Singh, Analysis of Repeated Measurement Data in the Clinical Trials, J. Ayurveda Integr. Med. 4 (2013), 77–81. https://doi.org/10.4103/0975-9476.113872.