Supplementary Material to: Tay CJX, Koh CWT, Qui M, et al. STAT: Enhancing confidence in clinical data analysis. Ann Acad Med Singap 2025;54:598-600. DOI: https://doi.org/10.47102/annals-acadmedsg.202570

Appendix S1. Methods.

Conceptualisation of STAT

The selection of statistical tools to be included in the STAT webtool was based on the most common measures necessary for healthcare professionals engaged in clinical research, such as observational studies, interventional studies and clinical audits.[1] These include descriptive statistics such as measures of central tendency and dispersion, measures of association such as relative risk and odds ratio, and common statistical tests such as Student's t-test, Chi-square test, ANOVA, Pearson's correlation and non-parametric equivalents.[1] Additionally, as data visualisation was a key area of need identified for clinical research but was not available in other freely available statistical software, we selected common graphing functions such as histogram, box-whisker plot, heatmap and scatterplot to be included in STAT.[2,3]

Data structure and case applications

The input dataframe for the STAT webtool is a structured, tabular dataset wherein each row represents a unique subject, and each column represents a variable of interest. Variables can either be continuous or categorical. Missing data should be left either as blanks or show "NA." Users are advised to pre-process their data to address missing values and ensure uniformity of values within columns before analysis. Most common examples of dataframes that can be analysed are demographic data, clinical features and laboratory results; thus, most types of healthcare data, such as clinical data, laboratory data and healthcare utilisation data, would be usable in STAT. Data requirements are aligned with other clinical data analysis tools such as IBM SPSS, GraphPad Prism and STATA.

STAT webtool features and operation

To use the STAT web tool, users can directly upload their research data in .csv or .xls formats. The data should be a single table of information, arranged in a wide format with 1 subject per row. The table should contain variable names in the first row, which will be automatically detected during upload. After uploading the data file, users can select the variables for downstream analyses. Subset analysis is also possible by filtering the data table using the variables prior to setting up the tests in the next section. To ensure research data security, all uploaded research data and all associated files are deleted immediately when the user exits the webtool. Users can also use the example dataset provided within the webtool to familiarise with the data format and the functionalities of the webtool.

STAT is divided into 4 sections arranged on different tabs: "Descriptive and normality testing", "Boxplot", "Comparing groups" and "Correlation" (Supplementary Fig. S1). The "Descriptive and Normality Testing" tab allows users to obtain descriptive statistics, including mean; standard deviation; and the 25th, 50th and 75th percentile values. For categorical variables, proportions of each category can be determined directly from STAT (Supplementary Fig. S1). To visualise the distributions of the variables, users can click on the widgets to display histograms or boxplots. Normality testing using the Shapiro-Wilks test is also possible, to evaluate if the distribution follows a Gaussian distribution, which will be helpful for statistical comparisons that require data to be normally distributed. Next, the "Boxplot" tab plots box and whisker charts according to categories selected from column headings in the uploaded dataframe (Supplementary Fig. S1). The "Comparing groups" tab allows users to perform statistical tests on any selected variables using Student's t-test, one-way ANOVA (with post-hoc testing), chi-squared test or their non-parametric equivalents. In addition, we have also included a relative risk and odds ratio calculator to allow users to directly evaluate relative risk (Supplementary Fig. S1). Lastly, the "Correlation"

tab allows comparison of 2–10 continuous variables using either Pearson or Spearman correlation, generating a correlation matrix with correlation coefficient and *P* value, a heatmap and a pair plot (Supplementary Fig. S1). As healthcare professionals may not be formally trained to perform statistical analyses and comparisons, we have also included brief annotations to guide users on the appropriate statistical tests.

Alternatively, users can deploy STAT locally, which will be useful when internet access is limited. All the files that are required to download are available on Github. STAT can also be hosted locally in a container on Docker Desktop, which could be used to run an instance of the webtool for subsequent repeated usage. Specific instructions on how to install and set up the STAT in the Docker Desktop app can be found in the README file in the Github repository.

Study conduct for assessment of user experience and confidence

Following the recruitment of participants, a total of 9 workshops were organised for groups of 5–20 individuals between 30 August 2023 and 11 June 2024, consisting of a 15-minute training session and a 30-minute practice data analysis. The practice data analysis had 3 clinical research questions (Supplementary Tables S2, S3 and S4) as a part of hands-on activity exercise during the workshop based on an example dataset (Supplementary Table S5). Subjects were requested to fill up the anonymous pre- and post-training questionnaire (Supplementary Appendix. User feedback questionnaire) hosted on Microsoft Forms platform. We collected demographics data, such as age and profession, data regarding prior non-clinical qualifications and experience with statistical analysis software, and user experience.

Statistical analysis

Statistical analyses were conducted using GraphPad Prism version 10.0.2 (GraphPad Software, LLC, Boston, US). Figures were created using GraphPad Prism version 10.0.2 and BioRender

(Science Suite Inc, Ontario, Canada). Statistical methods employed for post-study analysis includes two-tailed Mann–Whitney U-test for the comparison of unpaired continuous data between 2 groups, and Wilcoxon signed-rank test for the comparison of paired continuous data pre- and post-workshop. A *P* value of <0.05 was considered statistically significant.

REFERENCES

1. Krousel-Wood MA, Chambers RB, Muntner P. Clinicians' guide to statistics for medical practice and research: part I. Ochsner J 2006;6:68-83.

2. Ashour L. A review of user-friendly freely-available statistical analysis software for medical researchers and biostatisticians. Res Stat 2024;2:2322630.

3. MacDougall M, Cameron HS, Maxwell SRJ. Medical graduate views on statistical learning needs for clinical practice: a comprehensive survey. BMC Med Educ 2019;20:1.