BioData Mining

METHODOLOGY

Open Access

A simple guide to the use of Student's *t*-test, Mann-Whitney *U* test, Chi-squared test, and Kruskal-Wallis test in biostatistics

Davide Chicco[1,2*], Andrea Sichenze[1] and Giuseppe Jurman[3,4]

*Correspondence:
davide.chicco@unimib.it

[1] Università di Milano-Bicocca, Milan, Italy
[2] University of Toronto, Toronto, Ontario, Canada
[3] Humanitas University, Milan, Italy
[4] Fondazione Bruno Kessler, Trento, Italy

Abstract

In an age when machine learning and artificial intelligence are broadly employed, traditional statistics can still provide insightful information and results quickly and at a low computational cost. Statistics, in fact, offers many useful tools to researchers, including a series of univariate statistical tests that can identify relationships between pairs of numeric samples: Student's *t*-test, Mann-Whitney *U* test, Chi-squared test, and Kruskal-Wallis test. These tests generate several outcomes, including probability values (*p*-values) that can express a numerical quantity which accepts or rejects the null hypothesis, based on a certain threshold used. Although effective, these tests are often misused or employed in the wrong contexts, especially among biostatistics studies. Many scientific researchers do not seem to know how to choose one test over the others, and this misuse can lead to incorrect results and wrong conclusions. Here we present a simple theoretical and practical guide to the use of these four tests, first describing their theoretical properties and then displaying the results obtained by applying these tests to real-world medical datasets. Eventually, we explain when and how to use each test based on the data types of the samples considered. Our study can have a strong impact on scientific research by potentially influencing future studies involving these tests. Our recommendations, in turn, can help researchers produce more reliable and sound scientific results, thus increasing the quality of multiple scientific studies across various fields.

Keywords: Biostatistics, Chi-squared test, Kruskal-Wallis test, Mann-Whitney *U* test, Statistical significance, Student's *t*-test, Univariate statistical tests

Introduction

Today, machine learning and artificial intelligence have spread in all scientific fields, paving the way to a huge number of new scientific discoveries [1, 2]. These tools, even if useful, suffer from several flaws: their mathematical models often are complex and obscure, their application can be slow and computationally expensive, and their incorrect use can lead to several negative consequences [3, 4].

Biostatistics, on the contrary, offers several statistical tests that can serve to prove a relationship between variables, at a low computational cost, quickly, and in

human-understandable way. For these reasons, univariate statistical tests are broadly employed by researchers in several scientific fields nowadays. These statistical tests, however, sometimes are chosen in the wrong way or out of context, and this wrong selection can produce bad results and outcomes eventually. In this study, we try to get some order in this confused situation by providing a guide on four common univariate statistical tests: Student's *t*-test, Mann-Whitney *U* test, Chi-squared test, and Kruskal-Wallis test.

First, we provide an informal and purely explanatory definition of the statistical tests that will be analyzed later, accompanied by conceptual examples. The objective is to offer a general and straightforward understanding of each test.

The Student's *t*-statistic is a statistical procedure used to determine if there is a significant difference between two groups. For example, if we have two groups of patients and one receives drug A while the other receives drug B, we want to understand if there is a significant difference in their effects. There are three different types of *t*-tests: one-sample, independent samples, and paired samples. The one-sample *t*-test is used when we want to compare the mean of a sample to a known reference value. For instance, if a chocolate bar company claims that each bar weighs an average of fifty grams, we could take a sample of thirty bars that have an average weight of forty-eight grams. The one-sample *t*-test would help us determine if this average of forty-eight grams is significantly different from fifty grams. The independent samples *t*-test is applied when we want to compare the means of two independent samples to check if there is a significant difference between them. For example, if we want to test the effectiveness of two painkillers, A and B, we could use a sample of sixty people divided into two groups, one receiving drug A and the other receiving drug B. This type of *t*-test would allow us to assess if there is a significant difference in the effectiveness of the two drugs. The paired samples *t*-test is used when we want to compare the means of two dependent samples to see if there is a significant difference. For instance, to evaluate the effectiveness of a particular diet, we could weigh the same group of people before and after the diet and then observe the weight difference for each individual. It's important to highlight the difference between independent and paired samples: in the case of paired samples, the measurements are taken in pairs from the same individuals, such as results obtained from the same people before and after a treatment; in contrast, for independent samples, the individuals in the two groups are unrelated to each other. The one-sample test is very similar to the paired sample test, where we can think of paired samples as two samples taken at different times. We calculate the difference between paired values, obtaining a value and determining its significance based on how much this value deviates from a reference value [5, 6].

The Mann-Whitney test is a nonparametric test that checks the difference between two independent samples. For example: is there a significant difference in reaction times between men and women? The difference between the independent samples *t*-test and Mann-Whitney is that the *t*-test uses the value of the mean between the two groups, while Mann-Whitney *U* test uses the sum of ranks. To calculate the sum of ranks: order the subjects from lowest to highest value. The subject with the lowest value gets rank 1, the second-lowest rank 2, and so on, obtaining two groups with different ranks. The data

must therefore be rankable. The advantage of using the sum of ranks compared to the mean difference is that the data do not need to follow a normal distribution [7, 8].

The Chi-squared test is a statistical technique used to compare observed data with expected distributions based on a specific hypothesis. Its main purpose is to determine whether the differences between observed and expected values can be attributed to chance or if they suggest a significant relationship between variables. There are three common applications of the Chi-squared test: test of independence, goodness-of-fit test and test for homogeneity. The Chi-squared test of Independence is used to assess whether there is a relationship between two categorical variables. For example, researchers might want to investigate if gender influences the likelihood of having a Netflix subscription. In this case, a survey could be conducted to record whether individuals of different genders have a subscription. The data would then be arranged in a contingency table, allowing researchers to compare the observed frequencies to the expected frequencies calculated under the assumption that gender and subscription status are independent of each other. Another application is the Chi-squared goodness-of-fit test, which is used to evaluate if the observed frequencies for a categorical variable match the expected frequencies based on a known distribution. For instance, if a market researcher wants to find out whether the distribution of video streaming service subscriptions—such as Netflix, Amazon, and Disney—in a specific city aligns with the national distribution, this test would be appropriate. By comparing the observed data from the city with the expected frequencies derived from national statistics, the researcher can determine if the distribution significantly differs. In the end, the Chi-squared test for Homogeneity is used to check if different populations share the same distribution for a categorical variable. As an example, a survey might be conducted to explore if subscription rates for streaming services vary across different age groups. The observed data from each age group would be compared to determine if the distributions are similar or significantly different. The key differences among these tests lie in the number of variables involved and the research questions they address: the test of independence examines relationships between two variables, the goodness-of-fit test checks if a single variable matches a known distribution, and the test for homogeneity compares distributions across multiple groups In conclusion, the Chi-squared test proves to be a versatile tool for analyzing relationships between categorical variables, assessing distributions, and comparing different populations [9, 10].

The Kruskal-Wallis test is a non-parametric statistical method used to determine if there are significant differences between the medians of three or more independent groups. This test is particularly useful when the data do not follow a normal distribution, allowing researchers to make valid comparisons without relying on the assumptions required for parametric tests. For example, if a researcher who wants to find out if three different diets lead to different amounts of weight loss, the Kruskal-Wallis test can compare the median weight loss across these diet groups to see if the differences are statistically significant. This test is especially appropriate in situations where we have three or more independent groups to compare and when data are either ordinal or continuous but not normally distributed. It is also useful as a test of choice when we are interested in comparing medians rather than means, making it a flexible option for various types of data. For instance, a doctor might want to test if three different types

of physical therapy result in different recovery times for patients after surgery. By dividing the patients into three groups, each receiving a different therapy, the doctor could use the Kruskal-Wallis test to assess if there are significant differences in recovery times among these groups. The way the Kruskal-Wallis test works is straightforward yet effective. It begins by ranking all the data from the smallest to the largest, regardless of the group each data point belongs to. Then, it calcul ates the sum of these ranks for each group and checks if the sums differ significantly. If the test finds a significant difference, Kruskal-Wallis test suggests that at least one group is different from the others. In such cases, additional post-hoc tests can be performed to identify which specific groups differ. Overall, the Kruskal-Wallis test is a powerful tool for analyzing data that do not meet the assumptions required for parametric tests, making it a valuable technique in a wide range of research fields [11, 12].

Literature review

This section examines the four tests by reviewing their application in medical research using data acquired from EHRs. In the following text, we explain how each test has been employed, the type of data computed with the test and the results of the study. Of course, the list of articles is a subpart of all the studies involving the use of the four tests analyzed.

Student's t-test Daniel Moynihan and coauthors [13] conducted a study on the diagnostic processes for rare diseases, specifically Fabry Disease and Familial Hypercholesterolemia, using data from approximately one million patients across three hospitals in Singapore. A two-sample *t*-test for LDL-C levels was applied to compare confirmed cases (true positives) and suspect cohorts of FH, revealing no statistically significant difference.

Heather Patton et al. [14] conducted a study on a nonalcoholic fatty liver disease care pathway, evaluating adherence to components such as patient education, vibration controlled transient elastography examination, hepatology consultation, and weight management referral. They assessed the impact of the pathway on weight and ALT changes in 632 patients over a one-year period. The Student's *t*-test was used to analyze the impact of weight management participation and liver stiffness on weight change and ALT change, finding that referral to WM was associated with significant weight loss and ALT reduction

Cindy K. Blair and coauthors [15] used Student's *t*-test to compare Body Mass Index, tumor size, and age at diagnosis across groups defined by tumor characteristics (for example, stage, grade), finding significant associations between obesity and poor prognosis tumors.

Charlotte A. Nelson et al. [16] developed a method to predict chronic diseases, specifically multiple sclerosis, up to five years prior to clinical diagnosis by embedding electronic health record data into a biomedical knowledge graph and applying a random forest classifier. The Student's *t*-test was employed to compare the rank distributions of nodes within the SPOKE graph between patients with multiple sclerosis and patients without multiple sclerosis.

Heather Patton et al. [14] studied a care pathway for nonalcoholic fatty liver disease, assessing adherence to key steps and their effects on weight and ALT levels in over six hundred patients. The Student's *t*-test showed that referral to weight management was linked to significant reductions in both weight and ALT.

Cindy K. Blair et al. [15] used Student's *t*-test to compare body mass index, tumor size, and age by tumor traits, finding obesity significantly linked to poor-prognosis tumors.

Charlotte A. Nelson et al. [16] predicted multiple sclerosis up to five years before diagnosis using a biomedical knowledge graph and random forest. Student's *t*-test compared graph node ranks between patients with and without the disease.

Charles-Henri David and coauthors [17] conducted a retrospective study to assess the outcomes of patients with postcardiotomy cardiogenic shock supported with a left ventricular assist device on a 5-year period. A total of 29 patients (mean age 63) were included, with survival to discharge observed in more than half of cases. The mean duration of the support was 9 days, and 70% of patients were successfully weaned from the device.

Jing Gao and coauthors [18] studied the relationship between gut microbes, plasma trimethylamine N-oxide levels, and cardiovascular events in sixty participants, including patients with unstable angina, post-ST-segment elevation myocardial infarction, and healthy controls. Metagenomic sequencing and TMAO measurements revealed an association between specific gut microbial taxa and serum TMAO levels. A Student's *t*-test showed that elevated serum TMAO levels were linked to a higher risk of cardiovascular events, with certain microbial taxa acting as potential biomarkers for acute coronary syndrome diagnosis.

Rachel B. Issaka et al. [19] conducted a study to explore factors influencing colonoscopy completion after a positive fecal immunochemical test in a safety-net healthcare system. They analyzed data from over two thousand individuals aged between fifty and seventy-five who had an abnormal test result on a 3-year period. The study found that just over half of the patients completed their colonoscopy within one year of the positive test. The Student's *t*-test was used to compare the colonoscopy completion rates between different groups based on these factors, finding significant differences in the completion rates. The study highlighted the need for improved follow-up and documentation practices to ensure timely colonoscopy completion.

Andrew C. Storm and coauthors [20] conducted a study to assess the impact of implementing a new electronic health record system on patient safety and staff satisfaction in endoscopy suites. They compared procedure times and staff perceptions before and after the new system was introduced. The study found that procedure times significantly increased, and nurses spent less time directly monitoring patients after the implementation. The Student's *t*-test was used to analyze changes in procedure time and monitoring time, revealing significant differences.

Jiao-Zhi Zhou et al. [21] conducted a study on the associations between workplace bullying, burnout, and depression among clinical nurses in China, surveying 415 nurses across nine hospitals in October 2023. The study found that 20% of participants exhibited depression symptoms, with the depression group scoring significantly higher on the Negative Acts Questionnaire than the control group. The Student's

t-test was used to compare the depression group and the control group across variables such as age, length of service, and professional title.

Guangda Wang and coauthors [22] conducted a study involving 10 patients diagnosed with pre-malignant lesions and early-stage gastric cardia adenocarcinoma. The study found that in early-stage GCA, mutations in some of the genes were prevalent. The study also highlighted potential therapeutic targets based on genetic mutations, with 80% of patients showing promising treatment options. The Student's *t*-test was applied to compare the tumor mutation burden between different cohorts, showing no statistically significant differences between the HB cohort and the TCGA GCA cohort for most mutations; the test revealed significant differences in specific mutations, such as those in the EPHA2 gene.

Mann-Whitney U test Tam M. Do et al. [23] examined the relationship between dietary factors and breast cancer risk in Vietnamese women, using the Mann-Whitney *U* test to compare dietary intake and body mass index, between cases and controls. Significant differences were observed in the intake of vegetables, fruits, soybean products, coffee, and eggs, with higher consumption of these foods associated with lower breast cancer risk.

Brittany R. Lapin and coauthors [24] conducted a study involving over six thousand patients who completed patient-reported outcome measurements and satisfaction surveys at neurological clinics. The study found a significant positive association between PROM experience and visit satisfaction, with stronger effects observed among nonwhite patients, individuals with lower income, and those with more comorbidities. The Mann-Whitney *U* test was applied to compare PROMIS health scores and depression levels between patients who completed satisfaction surveys and those who did not, revealing significant differences in mental and physical health metrics between these groups.

Salvador Lugo-Perez et al. [25] conducted a study on rheumatoid arthritis patients without heart disease to investigate the relationship between antibody levels and left ventricular remodeling. The study revealed that patients with altered left ventricular geometry had significantly higher antibody titers, with anti-cyclic citrullinated peptide levels showing the strongest correlation with left ventricular remodeling. The Mann-Whitney *U* test was applied to compare echocardiographic variables and antibody levels between patients with and without left ventricular remodeling, highlighting statistically significant differences in these measures.

Ingo Steinbrück and coauthors [26] conducted a randomized controlled trial involving multiple centers to compare the safety and outcomes of cold versus hot endoscopic mucosal resection for non-pedunculated colorectal polyps. The study demonstrated that cold resection resulted in significantly fewer major adverse events, such as perforation and post-endoscopic bleeding, compared to hot resection. The Mann-Whitney *U* test was applied to compare continuous variables, including polyp diameter and resection speed, between the two groups, revealing significant differences in adverse event rates and procedural outcomes.

Ying Qian et al. [27] conducted a study to evaluate the effectiveness of copper bianstone scraping combined with a Chinese modified hypertension dietary therapy

program. The study included hundreds of hypertensive patients divided into a comparison group and an observation group. The results showed that the observation group had significantly greater improvements in blood pressure, body mass index, and blood glucose levels. The Mann-Whitney U test revealed significant differences between the groups, particularly in glycosylated hemoglobin, fasting blood glucose, and 2-hour postprandial blood glucose.

Jinghong Meng and coauthors [28] conducted a study on the incidence of surgical site infection in elective foot and ankle surgeries. More than a thousand of patients were included, and the 2% developed SSIs. The most common causative bacteria were pseudomonas aeruginosa and methicillin-resistant Staphylococcus aureus. Five factors were identified as independent risk factors for SSI: prolonged preoperative stay, use of allograft or bone substitute, elevated fasting blood glucose levels, low albumin levels, and abnormal neutrophil count. The Mann-Whitney U test was applied to data like age, hospital stay duration, and levels of albumin, fasting blood glucose, and other biological parameters, in order to compare the distributions between the groups with and without surgical site infection.

Brittany Lapin et al. [29] conducted a study to quantify the neurologic patient experience with patient-reported outcome measures across six neurology clinics. Over 16 thousands patients participated, completing both generic and condition-specific measures. The study identified that a majority of patients found the questions easy to understand, useful, and improved communication and control of their care. The Mann-Whitney U test was applied to compare the distributions of demographic factors, including age, income, and clinical characteristics, between groups of patients with varying experiences and satisfaction levels with measures.

Mitsuru Esaki and coauthors [30] conducted a study comparing endoscopic submucosal dissection outcomes in the postoperative stomach to those in the non-operative stomach for early gastric cancer, using propensity score matching to adjust for differences in baseline characteristics. Over one thousand patients were reviewed, with forty-one in the postoperative group and one thousand five in the non-operative group. The study found that the postoperative procedures required significantly more time but achieved similar rates of *en bloc*, complete, and curative resections with low adverse event rates. The Mann-Whitney U test was applied to compare procedure time and continuous variables like age, tumor size, and lesion depth.

Ausra Zelviene and Algirdas Bogusevičiu [31] conducted a study to evaluate the reliability and validity of the revised Champion's Health Belief Model Scale in measuring Lithuanian women's beliefs about breast cancer and screening, using data from three hundred fifty women aged forty to sixty-nine, with no prior mammograms or breast cancer history. The Mann-Whitney U test was used to compare the responses between women who participated in mammography screenings and those who did not. The test revealed that women participating in mammography screenings perceived lower susceptibility and severity but reported reduced confidence in performing breast self-examinations, while non-participants exhibited more barriers to mammography and lower perceived benefits. In bioinformatics, the Mann-Whitney U test has been used also for gene set enrichment analysis [32].

Chi-squared test Siyi Zhu et al. [33] conducted a study to assess the influence of obesity on the clinical and pathological characteristics of breast cancer and its impact on endocrine therapy effectiveness among Chinese patients. They analyzed data from nearly three thousand patients with luminal or human epidermal growth factor receptor two-negative early breast cancer, dividing them into obese and non-obese groups based on body mass index. The Chi-squared test was used to compare categorical variables, such as menopausal status, comorbidities, expression of estrogen and progesterone receptors, and luminal subtypes, between the two groups. Results indicated that obese patients had significantly higher rates of comorbidities and progesterone receptor positivity but showed no differences in other characteristics. This study highlights how obesity influences breast cancer features and endocrine therapy outcomes.

Di Zhao and coauthors [34] conducted a study to investigate the association between high body mass index and breast cancer characteristics across different age groups, utilizing Chi-squared tests to analyze relationships between body mass index and variables such as tumor grade and HER2 positivity. The results showed that in patients younger than fifty-five years, high body mass index was significantly associated with HER2 positivity and worse progression-free survival, while in patients older than fifty-five years, high body mass index was linked to lower tumor grades.

José Pablo Leone et al. [35] conducted an observational study using data from a large cancer registry to examine the effects of the COVID-19 pandemic on breast cancer diagnosis, treatment patterns, and short-term mortality. They analyzed over thirty-seven thousand cases of ductal carcinoma in situ and almost two hundred thousand cases of invasive breast cancer between 2018 and 2020. The Chi-squared test was used to evaluate differences in treatment patterns, specifically the distribution of surgery, chemotherapy, and radiation therapy across the years studied. The results showed that during 2020, there was a significant decline in breast cancer diagnoses, particularly for early-stage cancers. Treatment patterns also shifted, with reduced use of surgery and radiation and increased use of chemotherapy. However, the twelve-month mortality rates did not differ significantly between 2020 and the previous years.

The RECOVERY Collaborative Group [36] conducted a randomised, controlled trial to evaluate the safety and efficacy of convalescent plasma therapy for patients with COVID-19. Patients were randomly assigned to receive either usual care or usual care plus high-titre convalescent plasma. The primary outcome, 28-day mortality, was not significantly different between the two groups, with no substantial differences in other clinical outcomes such as hospital discharge or progression to invasive mechanical ventilation. A Chi-squared test was used to analyze observed effects in subgroups based on characteristics at randomization, including age, sex, and presence of anti-SARS-CoV-2 antibodies.

Juan Jin and coauthors [37] conducted a study to examine the characteristics of HER2-low breast cancer in comparison with HER2-zero and HER2-positive tumors, with a focus on clinical and molecular traits. The analysis included clinical and genomic data of five hundred seventy-nine metastatic breast cancer patients, and the HER2-low subtype was identified as more commonly associated with hormone receptor-positive tumors. The results indicated that, despite the differences in metastasis

patterns between HER2-low and HER2-positive patients in the hormone receptor-positive subgroup, the clinicopathological characteristics were largely influenced by hormone receptor status. Importantly, HER2-low tumors showed no significant differences in mutation alterations or copy number variations compared to HER2-zero tumors. Additionally, the study identified a higher prevalence of germline BRCA2 mutations in HER2-low patients. A Chi-squared test was employed to compare categorical variables, such as tumor characteristics and metastasis patterns, between the different HER2 subgroups.

Francesca Magnoni et al. [38] investigated the clinicopathological characteristics and outcomes of patients with pure and mixed invasive micropapillary breast cancer (IMPC) compared to invasive ductal cancer (IDC). The study included over thirty thousand IDC patients and more than two hundred IMPC patients. They found that both pure and mixed IMPC patients had a higher incidence of locally advanced disease and vascular invasion compared to IDC. After matching, pure IMPC patients showed worse overall survival and breast cancer-specific survival compared to IDC, whereas mixed IMPC patients had outcomes similar to IDC. The Chi-squared test was used to compare categorical variables, including tumor grade, lymph node status, vascular invasion, tumor molecular subtype, and type of surgery (mastectomy vs conservative surgery)

Kevser Tarı Selçuk and coauthors [39] conducted a study to determine the breast cancer screening behavior of women and investigate the relationship between health beliefs and screening behaviors. The study included 416 women aged forty and above. The results showed that the participation rates in breast self-examination, clinical breast examination, and mammography were relatively low. They found strong associations between perceived susceptibility, seriousness, self-efficacy, benefits, health motivation, and perceived barriers with screening behaviors. The Chi-squared test was used to compare screening behaviors with sociodemographic characteristics, and the variables considered in the analysis were age, education level, perceived economic status, and family history of breast cancer. The study also utilized multivariate binary logistic regression to examine how health beliefs influenced screening behaviors, and the Hosmer-Lemeshow test was applied to assess model fit.

Oindrila Bhattacharyya et al. [40] compared statewide health information exchange (HIE) data with survey self-report (SR) for measuring cancer screening. The study included Indiana residents and assessed screenings like colonoscopy, fecal immunochemical test, HPV and Pap tests, and mammography. Chi-squared tests were used to compare data on variables such as age, gender, and health status. Results showed fair to substantial concordance between HIE and SR, with higher sensitivity for procedures (for example, colonoscopy, mammography) and lower sensitivity for lab tests (for example, FIT, HPV). The Chi-squared test was used to evaluate differences in screening data.

Burcu Cengiz and coauthors [41] examined the effects of home-based nursing interventions, informed by the Health Belief Model, on individuals with stomas. The study included 30 participants in the experimental group and 31 in the control group. Data analysis involved Chi-squared, Mann-Whitney U, Wilcoxon, and Friedman tests to evaluate various outcomes. Results showed that home nursing interventions

significantly improved compliance with stoma management and reduced complication rates in the experimental group compared to the control group, but no significant difference was found in quality-of-life scores. The nursing interventions led to significant cost savings for the experimental group.

Yirong Wu et al. [42] developed prediction models using electronic health records to identify the "most harmful" breast cancers, focusing on variables like demographics, diagnoses, symptoms, procedures, medications, and laboratory results. They used Ridge Logistic Regression and Lasso Logistic Regression models, both of which showed strong performance with area under the ROC curve values of 0.818 and 0.839, respectively. Both models outperformed individual records components. The study also found that key features such as tobacco use and screening choices were associated with more harmful breast cancer cases.

Kruskal-Wallis test Jacob C. Clifton and coauthors [43] conducted a single-center retrospective study to evaluate the impact of an automated order reminder system on the timeliness of postanesthesia care unit order placement, analyzing over one hundred thousand surgical cases. Using the Kruskal-Wallis test, they assessed differences in procedural variables, including anesthesia duration, between patients receiving reminders and those who did not. The study found that reminders increased the likelihood of timely order placement and reduced delays in order submission, correlating with shorter PACU stays and improved pain management outcomes.

Mehrdad Karajizadeh et al. [44] conducted a study to identify the essential information needs for rapid response team electronic records in a major organ transplant center in Shiraz, Iran, through a cross-sectional survey of clinical roles. The Kruskal-Wallis test revealed significant differences among roles regarding data element priorities, and the Mann-Whitney U test highlighted distinctions specifically between registered anesthetist nurses and other groups.

Franklin Dexter and coauthors [45] analyzed workloads in veterinary dental clinics across forty-four workdays, using the Kruskal-Wallis test to compare anesthetist workloads among different days. They examined the workloads to identify significant variations in workload distribution and calculated upper confidence limits for workloads exceeding allocated times. The Kruskal-Wallis test was specifically applied to assess differences in workload data across various days, ensuring that variances in work schedules were adequately considered for accurate time allocation.

Irene L. Katzan et al. [46] conducted a survey to assess neurologic provider satisfaction with the systematic electronic collection of patient-reported outcome measures (PROMs), including disease-specific measures and depression screening using the Patient Health Questionnaire (PHQ-9). The survey was sent to two hundred ninety-nine staff physicians and advanced practice providers, with two hundred six responding. They used the Kruskal-Wallis test to evaluate differences in provider responses across age categories and provider types, and the Mann-Whitney U test to compare perceived usefulness between disease-specific information and the PHQ-9 depression screen. The results showed that PROM collection was deemed useful for patient care, research, and quality improvement, with respondents expressing similar perceptions of the clinical usefulness

of both disease-specific information and the PHQ-9 depression screen. The Kruskal-Wallis test revealed no significant differences in the usefulness between the two types of data within each center, suggesting that both were similarly valued by providers.

Daniel Clay Williams and coauthors [47] conducted a study to assess physician satisfaction with electronic health record (EHR) systems, particularly focusing on how they affect patient care. The study surveyed one hundred fifty-seven physicians at a quaternary care academic hospital. The results showed that overall satisfaction with the EHR system and its perceived impact on patient care were generally positive. The Kruskal-Wallis test was used for bivariate comparisons, and linear regression models identified that physician characteristics, including age and clinical role, were associated with EHR satisfaction. The perceived efficiency in using the EHR was found to be the most significant factor influencing overall satisfaction and the perception of its impact on patient care.

Karim Jabali et al. [48] conducted a study on anesthesiologists' perceptions of Electronic Health Records (EHRs) in Saudi Arabia, focusing on their impact, benefits, ease of use, and anesthesia-specific features. Sixty-seven anesthesiologists participated, with younger and more experienced users showing more favorable views. The study found that EHRs were seen as beneficial to healthcare delivery, though job rank negatively impacted perceptions of ease of use. Specific anesthesiology features were generally well-received, although integration challenges were noted. The Kruskal-Wallis test was applied to assess the influence of demographic factors on perceptions, revealing significant differences between age groups, but no differences based on job rank or experience in most aspects of EHR use. Statistical tests showed that age and EHR experience significantly influenced perceptions.

Martha Fors and coauthors [49] conducted a study to evaluate sex-dependent differences in the contribution of the neutrophil-to-lymphocyte ratio, platelet-to-lymphocyte ratio, lymphocyte-to-monocyte ratio, and mean platelet volume-to-platelet ratio to the severity and mortality of COVID-19 upon hospital admission. They analyzed data from 3,280 confirmed COVID-19 cases in Quito, Ecuador, and identified differences in the hematology ratios between men and women. Severe COVID-19 pneumonia and non-surviving patients had higher levels of these ratios, with men showing higher values than women. The Kruskal-Wallis test was used to assess the association between these ratios and mortality and severity, revealing significant differences between the sexes. Dunn's post hoc tests indicated differences in most groups, but no significant differences in NLR levels between surviving women and men. The results suggest that the hematology ratios could be useful biomarkers for predicting COVID-19 severity and mortality, with varying performances between men and women.

Luis E. Tollinche et al. [50] conducted a study on the implementation of an automated Clinical Alert System for improving the documentation compliance of the Immediate Preoperative Assessment (IPOA) at Memorial Sloan Kettering Cancer Center. A retrospective analysis of 42,285 cases, spanning a year before and after the alert system implementation, showed that the compliance rate increased significantly from seventy-six percent to eighty-eight percent ($p < 0.001$). The analysis revealed that compliance was more prominent in surgeries that were shorter in duration and occurred earlier in the day. The study suggests that automated alerts can effectively enhance adherence to

documentation standards, meeting regulatory requirements like those of the Centers for Medicare and Medicaid Services and The Joint Commission.

Sigal Shafran-Tikva and coauthors [51] conducted a study to investigate community-acquired pressure injuries (CAPIs) in older individuals using big data. Data were collected from over 44 thousand electronic medical records of patients admitted to internal medicine departments at two general medical centers between January 2016 and December 2018. The study found that the majority of CAPIs were located in the buttocks, with smaller occurrences in areas like the sacrum and ankle. Various tissue types, including necrotic and infected tissue, were identified. A significant portion of the CAPIs were second-degree, followed by first-degree and third-degree. The study revealed several patient characteristics associated with CAPIs, such as age, oxygen use, and albumin levels. Logistic regression was used to estimate the odds ratios of clinical variables for CAPIs, with a multivariate model highlighting variables that were clinically and statistically significant. The results underscore the importance of data-driven approaches in identifying and preventing CAPIs, particularly in elderly patients in community care settings.

This study This work is primarily intended for clinicians and biomedical researchers who do not have advanced statistical training and may lack the methodological tools to independently determine the appropriate use of these biostatistics univariate tests. Our aim is to equip these researchers with a structured and rigorous framework for selecting and distinguishing among the four tests examined, based on the nature of the data, the underlying research question, and key assumptions such as sample size and distributional properties

Even if some studies on guidelines to statistics exist [52–58], to the best of our knowledge no article provides a handbook to the four univariate statistical tests provided here. We fill this gap with our investigation. The objective of this study is to provide clarity on the correct use of these four tests by explaining when each can be applied, for which types of data, and under what conditions (specifically, sample size and normal distribution) they yield the most powerful results. Specifically, we aim to help researchers select the most appropriate test to detect differences in means between groups or to assess the association between categorical variables.

We organize the rest of the manuscript as follows. After this Introduction, we explain the mathematical properties of the four tests in the Methods section, where we also offer guidance on which univariate statistical test to select based on the data types of the samples analyzed. Afterwards, we report and describe the results of the application of these tests on toy artificial data and on real-world medical data in the "Results" section. Eventually, we summarize and discuss the main findings of our study and describe some limitations and future directions in the "Discussion and conclusions" section.

Methods

In this section, we introduce the formal definitions, mathematical properties, and assumptions underlying the statistical tests used in this analysis. We also discuss the calculation and interpretation of the p-value, as well as the conditions required for

each test to be valid. Additionally, we examine the tests for normality, which help determine whether parametric or non-parametric tests are more appropriate. This information will allow us to better understand the strengths and limitations of each test in the context of the datasets analyzed.

The null hypothesis is the uninteresting scenario

All the four tests produce probability values (called p-values) that can be used to confirm or reject the null hypothesis (denoted H_0 or h_0). A null hypothesis is a statement which says there is no effect or no difference between two samples or events, and it serves as a starting point for statistical testing.

We clarify the meaning of this concept through an example. Let us consider the scenario where a pharmaceutical company wants to test whether a new medication is effective in lowering cholesterol compared to a placebo. In this scenario, the null hypothesis would be that the new medication has no effect on cholesterol compared to the placebo.

In statistical terms, this can be expressed as: H_0: $\mu_{drug} = \mu_{placebo}$

where μ_{drug} represents the mean cholesterol level of the group taking the new drug and $\mu_{placebo}$ the mean cholesterol level of the group taking the placebo. In this study, the researchers would collect cholesterol data from two groups (one receiving the new drug and the other receiving the placebo) and perform a statistical test (one of the four tests described in this study) to determine if there is enough evidence to reject the null hypothesis in favor of an alternative hypothesis, which would suggest that the new drug does have an effect on lowering cholesterol. That means, if the p-value produced by the statistical test applied to the cholesterol levels of the group taking the new drug (μ_{drug}) and the cholesterol levels of the group taking the placebo ($\mu_{placebo}$) is lower than the significance 0.005 threshold [59], we can reject the hypothesis that the new medical has no effect on cholesterol compared to the placebo.

In pseudo-code, we can write the previous claim this way:

```
p_value ← test(drug_group.cholesterol_levels, placebo_group.cholesterol_levels)
if(p_value < 0.005) then { null_hypothesis ← false }
```

In turn, this statement means that we can accept the alternative hypothesis saying that there the new medicine has some effect on lowering cholesterol compared to the placebo. On the contrary, if the p-value were higher than the 0.005 threshold, we could not reject the null hypothesis (we will explain significance thresholds later on more precisely).

One can think of null hypotheses as uninteresting, irrelevant, normal scenarios [60].

Properties and theoretical aspects of statistical tests

In this section, we explain the theory beyond the four statistical tests analyzed in this study.

Student's t-test

The Student's t-test is a statistical test used to verify the statistical significance of the difference in responses between two groups. It is based on the Student's t-distribution, used with small sample sizes and a reference population with a normal (or approximately normal) distribution. The test compares the means of two groups of data and evaluates whether the difference between them is statistically significant relative to the variability within the groups themselves.

The test provides a t-value and an associated p-value: the t-value represents how large the difference between the means of the two groups is relative to the variability within the groups; the p-value represents the probability that the observed difference is due to chance, assuming there is no real difference between the group means. If the p-value is low (in our case, less than 0.005), it indicates that the difference between the means is statistically significant.

Assumptions:

- One-sample t-test: requires one sample and a reference value.
- Independent samples t-test: requires two independent samples. The variances must be approximately equal, which can be checked using Levene's test [61].
- Paired samples t-test: requires paired samples.

We can evaluate differences only on metrics, not on categorical values.

Hypotheses:

- One-sample t-test: Null hypothesis: the sample mean is equal to the reference value.
- Independent samples t-test: Null hypothesis: the means of the two samples are equal.
- Paired samples t-test: Null hypothesis: the mean difference between paired groups is zero.

How to calculate the t-value:

$$t = \frac{\text{difference between sample means}}{\text{standard error (standard deviation from the mean)}}$$

- One-sample t-test:

$$t = \frac{X - \mu}{S/\sqrt{n}}$$

where X is the sample mean, μ is the known mean, S is the standard deviation of the data, and n is the sample size. We recall that, in the previous formula, the denominator division is computed before the general division.

- Independent samples t-test:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

where $\bar{X}_i$ is the sample mean, s_i is the standard deviation, and n_i is the sample size for group i.

- Paired samples t-test:

$$t = \frac{\bar{X}}{S/\sqrt{n}}$$

where $\bar{X}$ is the mean of the differences, S is the standard deviation of the differences, and n is the number of pairs.

Results of the t-statistic: The higher the t-value, the greater the difference between the means of the samples (the same applies in reverse). The greater the dispersion of the mean, that is, the standard error, the less significant the difference in the means between the data.

To accept or reject the null hypothesis, we can either consult a critical t-value table or calculate the p-value.

The p-value is the probability of observing the experimental data or the test results, assuming the null hypothesis is true. If the p-value is smaller than a chosen threshold (in our case, 0.005), it means we have sufficient evidence to reject the null hypothesis and conclude that there is a difference between the two groups.

Regarding the t-value table, we choose a significance level (0.005), then look at the value for $1 - 0.005 = 0.995$. Next, we choose the degrees of freedom (df). For one-sample and paired tests, $df = n - 1$, while for independent samples, $df = n_1 + n_2 - 2$.

If the calculated t-value is greater than the critical t-value, we reject the null hypothesis.

We also need to consider the difference between direct and indirect hypotheses. In an indirect hypothesis, if we verify that there is a difference between the means of two groups, we do not know which group has a higher mean than the other. In a direct hypothesis, we know this information (Table 1).

For these different types of the Student's t-test, it is possible to calculate the values that the t-statistic tends toward as the variables change (boundary values). Recall the definition of the variables involved in the calculation:

- $\bar{X}$: sample mean

Table 1 Differentiated Student's t-tests. df: degrees of freedom. The degree of freedom in statistical tests refers to the number of independent values or quantities that can vary in an analysis without breaking any constraints

Sample	Difference	Hypothesis	Formula	DF
one sample	one sample with a known reference value μ	h_0: $\mu = \bar{x}$, h_1: $\mu \neq \bar{x}$	$t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$	$n-1$
independent samples	two independent samples with similar variances	h_0: $\bar{x}_1 = \bar{x}_2$, h_1: $\bar{x}_1 \neq \bar{x}_2$	$t = \frac{\bar{x}_1-\bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)}}$	n_1+n_2-2
paired samples	differences in values from the same sample at different times δ	h_0: $\delta = 0$, h_1: $\delta \neq 0$	$t = \frac{\bar{x}}{s/\sqrt{n}}$	$n-1$

- μ: known mean (only for the one-sample t-test)
- S: standard deviation
- n: sample size

In general, we note the following common limits:

- $S \to 0$: $t \to \infty$: a very small standard deviation erroneously suggests strong statistical significance.
- $n \to \infty$: very large samples make even very small differences significant.
- $n \to 0$: very small samples make it difficult to obtain significance even with large differences.

Mann-Whitney U test

The Mann-Whitney U test, also known as the Wilcoxon-Mann-Whitney test, is a non-parametric statistical test used to evaluate the significance of differences between two independent groups. Unlike the Student's t-test, it does not require the assumption of normality in the underlying population distributions, making it suitable for data that may not follow a normal distribution or when sample sizes are small.

The test compares the ranks of the data rather than their raw values, assessing whether one group tends to have systematically higher or lower ranks than the other. It is particularly effective for ordinal or continuous data that violate parametric assumptions.

The test is based on the following assumptions:

- The two samples are independent.
- Observations are ordinal or continuous.
- The distributions of the two groups are similar in shape; otherwise, the test may capture differences in shape or dispersion in addition to central tendency.

The hypotheses for the Mann-Whitney U test are:

- Null hypothesis (H_0): The distributions of the two groups are identical, implying no systematic difference in ranks.
- Alternative hypothesis (H_1): The distributions of the two groups differ, indicating a systematic rank difference between groups.

The test produces the U statistic, which represents the sum of ranks assigned to one of the groups, adjusted for the sample sizes. This statistic is then converted into a p-value to assess significance.

The test statistic U is computed through the following formula:

$$U = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1 + 1)}{2} - R_1$$

where:

- n_1, n_2 are the sample sizes of the two groups.

- R_1 is the sum of ranks in the first group [7].

Regarding interpretation, it is worth noticing that a smaller U value indicates a larger difference between groups. Additionally, the p-value derived from U determines whether the observed rank differences are statistically significant. A p-value below the significance level (for instance, 0.005) leads to rejecting H_0, supporting that the groups differ significantly.

This test has several key features: robustness, since the test does not assume equal variances or normality in the data; applicability, since this test is suitable for both small and large sample sizes but requires independent observations between groups; and flexibility, since this test can handle ordinal data and non-normal continuous data [62].

The U test presents several advantages. It is applicable to skewed data or data with outliers, it is Suitable for small sample sizes, and it does not assume homogeneity of variances.

However, this test has some drawbacks as well: it is Less powerful than the t-test for normal distributions, and it may lose sensitivity when data contains ties or ranks overlap significantly [63].

Chi-squared test

The Chi-squared test is a statistical method used to evaluate the differences between observed and expected frequencies in categorical data. It was first introduced by Karl Pearson in 1900 as a measure of discrepancy between observed and theoretical distributions. The test is widely applied in contingency table analysis and goodness-of-fit testing [64].

The Chi-squared statistic is calculated as:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

where:

- O_i represents the observed frequency in the i-th category,
- E_i represents the expected frequency in the i-th category under the null hypothesis [9].

The Chi-squared test has several relevant characteristics. First, it evaluates whether the observed frequencies significantly deviate from the expected frequencies. The null hypothesis (H_0) states that the observed frequencies match the expected frequencies. Moreover, the alternative hypothesis (H_1) states that the observed frequencies differ from the expected frequencies.

This test makes some assumptions as well. Expected frequencies should be sufficiently large, typically at least 5 in each category, to ensure the validity of the test. Observations must be independent, meaning that each data point should not influence or depend on others. Data must be derived from random samples representing the population of interest.

The Chi-squared statistic follows a Chi-squared distribution with $k-1$ degrees of freedom, where k is the number of categories. The critical value of χ^2 depends on the chosen significance threshold (for instance, $\alpha = 0.005$) and the degrees of freedom.

Regarding interpretation, we can state that a high χ^2 value indicates a large difference between observed and expected frequencies. Additionally If the p-value associated with χ^2 is less than the significance level, the null hypothesis is rejected, suggesting significant differences. Conversely, if the p-value is greater than the significance level, there is insufficient evidence to reject the null hypothesis.

This Chi-squared test suffers from some limitations, too. When $O_i = E_i$ for all i, $\chi^2 = 0$, indicating no difference between observed and expected frequencies. As $(O_i - E_i) \to \infty$, $\chi^2 \to \infty$, reflecting increasing discrepancies. If E_i values are too small, the Chi-squared statistic may not follow the theoretical Chi-squared distribution, compromising the test's reliability [9, 64].

Kruskal-Wallis test

The Kruskal-Wallis test, also called Kruskal-Wallis H-test sometimes, is a nonparametric statistical procedure used to assess whether there are significant differences among three or more independent groups based on ordinal or continuous dependent data. It is often considered an extension of the Mann-Whitney U test, which is limited to comparing two groups. The Kruskal-Wallis test evaluates whether the distributions of ranks within groups are statistically equivalent or if at least one group deviates significantly from the others; this test is particularly useful for comparing independent groups when the assumptions of one-way ANOVA test are not met, such as when data are non-normal or when sample sizes are unequal [11].

The test operates on the following principles:

- Observations within each group are ranked across the entire dataset, assigning rank 1 to the smallest value, rank 2 to the next smallest, and so on.
- The test statistic H is computed using the sum of ranks for each group and measures the degree to which the ranks differ among groups.
- The null hypothesis (H_0) states that all independent groups share the same central tendency, indicating they come from the same population distribution.
- The alternative hypothesis (H_1) posits that at least one group exhibits a different central tendency, suggesting it originates from a distinct population distribution.

Assumptions of the Kruskal-Wallis test include:

- Observations within groups must be independent.
- Groups must have distributions with the same shape, though not necessarily the same median.
- Data must be at least ordinal and originate from random independent samples.

The test statistic H is calculated as:

$$H = \frac{(n-1)}{12} \sum_{i=1}^{k} \frac{n_i \left(\bar{R}_i - E_R\right)^2}{\sigma_R^2}$$

where:

- n is the total number of observations,
- k is the number of groups,
- n_i is the sample size of group i,
- $\bar{R}_i$ is the average rank for group i,
- E_R is the expected rank under the null hypothesis, given by $E_R = \frac{n+1}{2}$,
- σ_R^2 is the rank variance, calculated as $\sigma_R^2 = \frac{n^2-1}{12}$.

Degrees of freedom for the test are defined as $\mathrm{df} = k - 1$. The computed H statistic is compared against the critical value from the Chi-squared distribution table with df degrees of freedom at the desired significance level.

Regarding interpretation, if the H statistic exceeds the critical value, or if the p-value is less than α, the null hypothesis is rejected, indicating significant differences among the groups. If H is below the critical value, or the p-value is greater than α, there is insufficient evidence to reject the null hypothesis.

Of course, the H statistic should not be confused with the null hypothesis (usually indicated as H_0 or h_0).

This Kruskal-Wallis test also presents several limitations. When the distributions of all groups are identical, the test statistic H is close to zero, indicating no difference in medians. As the rank differences between groups increase, $H \to \infty$, reflecting stronger evidence against the null hypothesis [65]. However, when sample sizes are small or unequal, or when there are many tied ranks, the approximation of H to a chi-squared distribution becomes less accurate, potentially reducing the test's reliability [62].

Python software packages

In our computational analyses, we utilized several Python packages to handle data manipulation, statistical testing, and visualization:

- `Pandas`: for data manipulation and analysis (reading datasets, handling dataframes, and cleaning data) [66];
- `SciPy`: for Student's t, Mann-Whitney U and Chi-squared tests [67];
- `Matplotlib`: to create bar charts displaying the p-values from statistical tests [68];
- `NumPy`: for normality testing (Shapiro-Wilk test and Kolmogorov-Smirnov test) [69].

We decided to use the Python programming language because it is free and open source, allowing the universal reproducibility of our experiments [70, 71]. Additionally, Python is considered the most widely used programming language in the world, by the PYPL index of July 2025 [72], the TIOBE index of July 2025 [73] and the 2022 Kaggle survey [74].

Table 2 Summary of statistical tests analyzed. n_1 and n_2 are the sample sizes of the two groups in the comparison. Normality required: Shapiro-Wilk test significant for $n < 50$ or Kolmogorov-Smirnov test significant for $n \geq 50$

Test	Data types for both samples	Comparison	Normality required	Output	purpose
Student's t	numerical	means of 2 groups	yes	$t \in (-\infty; +\infty)$	differences between group means
Mann-Whitney U	numerical, ordinal, or categorical	ranks of 2 groups	no	$U \in [0; n_1 \cdot n_2]$	compare ranks between groups. n_1 and n_2 are the sample sizes of the two groups in the comparison
Chi-squared	categorical	observed versus expected frequencies	no	$\chi^2 \in [0; +\infty)$	association or independence between variables
Kruskal-Wallis	numerical, ordinal, or categorical	ranks of ≥ 2 groups	no	$H \in [0; +\infty)$	compare ranks across multiple groups

Table 3 Assumptions and limitations of the statistical tests considered in this study

Test	Assumptions	Limitations
Student's t	normally distributed data, equal variances across groups	sensitive to outliers
Mann-Whitney U	independent observations, data not necessarily normal	does not provide direct information on the difference between groups, less powerful than the Student's t-test
Chi-squared	data not necessarily normal, independent observations	does not provide information about the direction or magnitude of the difference
Kruskal-Wallis	independent observations	does not provide information about the direction or magnitude of the difference

Summary of the four tests considered

We can recap the meanings of the four statistical tests considered this way. The Student's t-statistic measures the standardized difference between two group means: a value of zero indicates no difference, while larger positive or negative values suggest greater separation. The Mann-Whitney U statistic compares the ranks of values from two independent groups: values near zero indicate minimal difference, and values approaching the product of the group sizes indicate strong ordering in favor of one group. The Chi-squared statistic compares observed and expected frequencies in contingency tables; it starts at zero when distributions match perfectly and increases with greater discrepancies. Similarly, the Kruskal-Wallis H statistic evaluates whether rank distributions differ across multiple groups, starting at zero when distributions are identical and increasing with larger differences. Higher values of these statistics generally provide stronger evidence against the null hypothesis, but their significance must be evaluated using corresponding p-values and degrees of freedom: these values are not p-values, but rather the raw statistics calculated during the test. We recap the properties of the four analyzed statistical tests in Table 2 and their assumptions and limitations in Table 3.

The four tests produce not only a probability value (p-value), but also other values. As we explained earlier ("Properties and theoretical aspects of statistical tests" section), the

Table 4 Lower and upper boundary cases for the considered statistical tests. n_1 and n_2 are the sample sizes of the two groups in the comparison. *H*: statistic produced by the Kruskal-Wallis test, that indicates the degree of difference among the groups being compared, and should not be confused with the null hypothesis (H_0 or h_0)

Test	Lower limit	Upper limit
Student's *t*	0 = no difference between groups	$t \to +\infty$ or $t \to -\infty$: larger differences between means
Mann-Whitney *U*	0 = no difference between groups	$U = n_1 \cdot n_2$: maximum possible comparisons
Chi-squared	0 = no difference between groups	no defined upper limit; χ^2 increases with discrepancies
Kruskal-Wallis	0 = no difference between groups	no defined upper limit; *H* increases with discrepancies

Table 5 Our interpretation of the *p*-values

P-value range	Judgment	Interpretation
[0; 0.005)	small	strong statistical evidence against H_0; a very significant effect or difference between groups.
[0.005; 0.01)	moderate	some statistical evidence against H_0; some significant effect or difference between groups.
[0.01; 0.05)	large	limited statistical evidence against H_0; marginally significant effect or difference between groups.
[0.05; 1)	very large	no statistical evidence against H_0; the observed effect or difference might be due to chance.

Student's *t*-test generate the *t*-value, the Mann-Whitney *U* test generate the *U* statistic, the Chi-squared test delivers the Chi-squared statistic, and the Kruskal-Wallis test outputs the *H* statistic. We report the lower limits and the upper limits of these statistics in Table 4.

Even if these statistics can be useful in several contexts, we mainly consider these four statistical tests for the probability values they produce. The *p*-value is a statistical measure that quantifies the probability of observing an effect or a difference as extreme as, or more extreme than, what is obtained from the data, assuming the null hypothesis (H_0) is true. In other words, it is an indicator of the strength of evidence against H_0.

To summarize:

- A small *p*-value indicates that the observed data are unlikely under H_0, suggesting the null hypothesis should be rejected in favor of the alternative hypothesis (H_1).
- A large *p*-value suggests there is insufficient evidence to reject H_0, but it does not necessarily confirm its validity [75].

In our study we will use the significance threshold at $\alpha = 0.005$, that actually is 5×10^{-3}, as suggested by Daniel J. Benjamin et al. [59]: this choice is made to improve the reproducibility of scientific research and reduce the false positive rates in any scientific field (Table 5).

Misuses of p-values Probability values produced by statistical tests are the cornerstone of thousand of scientific studies worldwide, but their use can sometimes be

wrong or misleading. In particular, the meaning of resulting p-values can be misunderstood [76–78] or the threshold to decide if a p-value is significant or not can be arbitrarily chosen [79–81].

Although widely used across scientific disciplines, p-values are often misunderstood and misapplied. A common misconception is that a p-value indicates the probability that the null hypothesis is true, when in fact it represents the probability of obtaining results at least as extreme as the observed ones under the assumption that the null hypothesis is correct [76–78]. This subtle but crucial distinction is frequently overlooked, leading to erroneous conclusions, such as equating $p > 0.05$ with evidence of no effect, or interpreting $p < 0.05$ as confirmation of the alternative hypothesis [77].

The threshold of $p < 0.05$, though historically influential, is arbitrary and can create a false dichotomy between "significant" and "non-significant" results [78, 79]. As Altman et al. [76] and Di Leo et al. [79] point out, this practice not only ignores the continuity of evidence but also discourages nuanced interpretations based on confidence intervals, effect sizes, and prior plausibility. Even when $p < 0.05$, the evidence may be weak—for instance, a p-value of 0.05 corresponds to a Bayes factor upper bound suggesting the alternative hypothesis is only about 2.5 times more likely than the null [76].

Moreover, the fixation on arbitrary significance thresholds fosters questionable research practices such as p-hacking, where analytical flexibility is exploited to reach statistical significance [81]. Evidence of widespread p-hacking has been documented through the analysis of p-curve distributions, revealing a disproportionate number of results just below the 0.05 threshold, particularly in biomedical and multidisciplinary research [81].

In response to these issues, the American Statistical Association and numerous scholars have called for abandoning the binary notion of *statistical significance* [80]: their members advocate instead for more informative inferential approaches that emphasize compatibility intervals, transparent reporting, prior evidence, and contextual interpretation of results. Lowering the significance threshold (for example, to 0.005) or integrating Bayesian reasoning and estimates of false discovery rates can also help mitigate the risks of spurious findings [76, 79]. Ultimately, the meaningful use of p-values demands a shift from threshold-based decisions to comprehensive inferential reasoning, where uncertainty, effect size, study design, and scientific plausibility are given equal weight in interpreting results.

Selection of statistical tests

These tests are employed depending on the nature of the variables under consideration. The following table summarizes the appropriate choice of test based on the type of comparison between the variables. In the following sections, n will indicate the sample size and k the number of groups in the comparison, which is the number of unique elements in a data samples. For example, a sample containing the $\{1, 2, 3, 4, 5, 5, 5\}$ data elements has $n = 7$ (which is the size of the sample) and $k = 5$ (which is the number of unique data elements of the sample).

In this section, we set the record straight about which statistical tests should be used in which cases. We summarize all our guidelines in Table 6, Figs. 1, and 2.

Table 6 List of statistics tests to be used for comparing pairs of samples, depending on the data type of the samples and on the number of distinct data elements of the samples. Comparison of statistical tests based on different types of variables, sample size n, and number of groups k. Group: number of unique data elements in a sample (k in the text). The case numbers, Diamond 1, and Diamond 2 refer to Figs. 1 and to 2. *Even when the one-way ANOVA test assumptions are valid, using the Kruskal-Wallis test is still correct, even if less powerful than the one-way ANOVA test

Case number	Groups	Comparison	Recommended test
1	$= 2$	ordinal versus ordinal	Chi-squared test
2	$= 2$	categorical versus categorical	Chi-squared test
3	$= 2$	categorical versus ordinal	Chi-squared test
4	$= 2$	numeric versus numeric	if $n < 50$ and Shapiro-Wilk test is significant or $n \geq 50$ and Kolmogorov-Smirnov test is significant (Diamond 1):
			Student's t-test (only with normal distribution, more powerful) else:
			Mann-Whitney U test (for any distribution, including normal)
5	$= 2$	numeric versus categorical	Mann-Whitney U test
6	$= 2$	ordinal versus numeric	Mann-Whitney U test
7	≥ 3	numeric versus numeric	if data follows a normal distribution & homogeneity of variances (Diamond 2): one-way ANOVA test (more powerful) else:
			Kruskal-Wallis test*
8	≥ 3	numeric versus categorical	Kruskal-Wallis test
9	≥ 3	categorical versus categorical	Kruskal-Wallis test
10	≥ 3	categorical versus ordinal	Kruskal-Wallis test
11	≥ 3	ordinal versus numeric	Kruskal-Wallis test
12	≥ 3	ordinal versus ordinal	Kruskal-Wallis test

Comparison for $k = 2$ groups. When the pair of samples contains only two different data elements, all the tests' selections are straightforward except the numeric versus numeric comparison. The Chi-squared test should be employed for the ordindal-ordinal case, the categorical-categorical case, and the categorical-ordinal case. The Mann-Whitney U test should be utilized for the numeric-categorical and the ordinal-numeric test (Table 6 and Fig. 1).

The situation is different for the numeric-numeric statistical comparison, where two tests can be utilized. The choice between the Student's t-test and the Mann-Whitney U test is primarily based on the assumption of normality of the data. If the data follow a normal distribution, the Student's t-test should be preferred. However, if the data do not meet this assumption, the Mann-Whitney U test is the appropriate test to use. While the Mann-Whitney test can be employed even with normally distributed data, it may result in a loss of statistical power, which could impair the detection of differences between the groups.

The Student's t-test and the Mann-Whitney U test have some differences in calculation. The Student's t-test calculates the difference between the means of the two groups and compares it with the variability (variance) within each group. The formula takes into account the sample size n and the variance of both groups. On the other hand, in the Mann-Whitney U test, instead of working with raw numerical values,

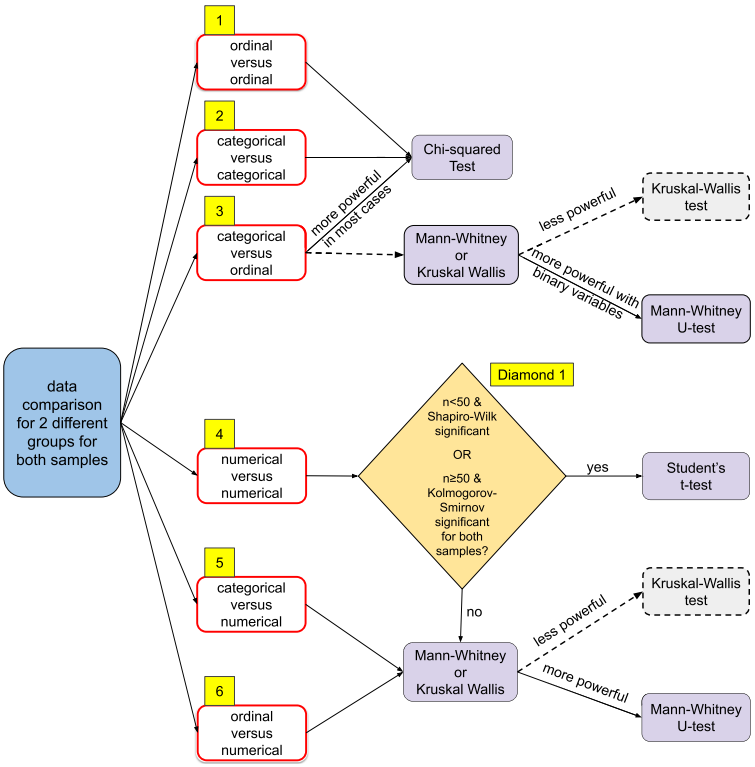

Fig. 1 Flowchart all possible statistical tests choices for pairs of samples containing data of 2 distinct groups. The numbers in the yellow boxes correspond to the case numbers in Table 6. Groups: number of unique elements in a sample

the data are ranked from smallest to largest. The sum of ranks for each group is computed, and the ranks (rather than the original data points) are compared between the groups.

Difference raise also when dealing with outliers. The Student's t-test is sensitive to outliers, as it is based on the mean of the data, which can be heavily influenced by extreme values. In contrast, the Mann-Whitney U test is less sensitive to outliers, as it operates on ranks rather than raw values.

We recap our pieces of advice in Table 6 and Fig. 1.

Comparison for $k \geq 3$ groups. In cases where three groups or more need to be compared, the choice of the statistical test is between one-way ANOVA test and Kruskal-Wallis test. The one-way ANOVA test is preferred when the assumptions of normality, homogeneity of variances, and independence are met, providing slightly higher statistical power compared to the Kruskal-Wallis test. However, studies have shown that the difference in power between one-way ANOVA test and Kruskal-Wallis test is minimal under normality conditions, with Kruskal-Wallis presenting a disadvantage of approximately 0.01 to 0.02 in power. When the assumptions for one-way ANOVA test are violated, the Kruskal-Wallis test often demonstrates substantially higher power, especially under non-normal distributions such as the Chi-squared distribution with $df = 2$. Given that perfect normality is rarely achieved in practice, the Kruskal-Wallis test offers

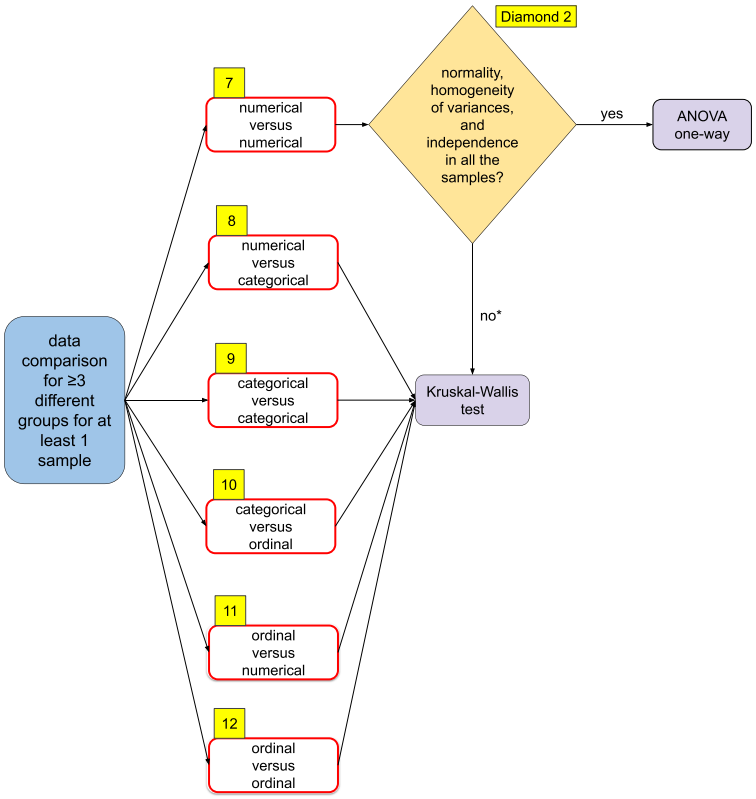

Fig. 2 Flowchart representing all possible statistical tests choices for pairs of samples containing data of 3 distinct groups or more. The numbers in the yellow boxes correspond to the case numbers in Table 6. Groups: number of unique elements in a sample. *Actually, even when normality, homogeneity, and independence in both samples are present, the Kruskal-Wallis test is still valid and has been proved by Jamie Gleasonto be slightly less powerful than the one-way ANOVA test [82]

a robust alternative with little to lose in terms of power, making it a reliable choice for non-normally distributed data or when assumptions of homogeneity of variances are in question [82]. We report our recommendations in Table 6 and Fig. 2.

Since the strict conditions for a proper use of the one-way ANOVA test are rarely met among biomedical data, we decided not to focus on this particular test in this study. We mention the existence of the one-way ANOVA test for full disclosure and for clarity purposes.

Normality testing

It is essential to confirm the normality of the data before selecting the appropriate test. Two common normality tests used for this purpose are the Shapiro-Wilk test [83] and the Kolmogorov-Smirnov test [84]. The condition of normality is crucial in deciding which test to use, and to have statistical certainty about the normality of the data, these two tests are commonly employed.

The Shapiro-Wilk test is one of the most common tests used to check if a dataset follows a normal distribution. It is particularly suitable when you have a small or medium-sized sample (up to about two thousand observations). How it works:

- Null hypothesis (H_0): the data follows a normal distribution.
- Alternative hypothesis (H_1): the data does not follow a normal distribution.

Procedure:

- Test statistics are calculated based on the difference between your observed data and the values you would expect if the data were normally distributed.
- The test returns a p-value.
- If the p-value is greater than the significance level, you do not reject the null hypothesis, meaning the data are considered compatible with a normal distribution.
- If the p-value is less than the significance level, you reject the null hypothesis and conclude that the data do not follow a normal distribution [83].

Limitations of the test:

- Sensitivity to large samples: with very large samples, even small deviations from normality may be statistically significant, even if these deviations are not practically significant.
- Sensitivity to small samples: with very small samples, the test may not have enough power to detect deviations from normality [85].

When to use it: The Shapiro-Wilk test is useful to determine whether you can use parametric tests (which require normality, like the t-test or one-way ANOVA test) or non-parametric tests (like the Mann-Whitney or Kruskal-Wallis test). Note: Since this is a test for normality, it is only valid for continuous numeric variables and not applicable to ordinal or categorical variables.

The Shapiro-Wilk test is generally preferred over the Kolmogorov-Smirnov test for assessing the normality of data in specific situations:

- Higher statistical power for small samples: The Shapiro-Wilk test is known to be more powerful in detecting deviations from normality when working with small samples (typically fewer than 50 observations), whereas the Kolmogorov-Smirnov test may not be as sensitive in these situations.
- Test specifically for normality: The Kolmogorov-Smirnov test compares an empirical distribution with a reference distribution (which could be normal or another), whereas the Shapiro-Wilk test is specifically designed to test whether the data follow a normal distribution. Therefore, the Shapiro-Wilk test is more targeted when the goal is to assess normality.
- Asymptotic distribution: The Kolmogorov-Smirnov test is a non-parametric test and may be more suitable for large samples, but in the case of small samples, the asymptotic distribution it is based on may not provide results as accurate as those from the Shapiro-Wilk test [86].

In summary, the Shapiro-Wilk test is preferable for small samples and when the goal is solely to check normality, while the Kolmogorov-Smirnov test is better suited for

comparing an empirical distribution with a theoretical distribution, especially with larger samples.

We report the use of these tests in the decision Diamond 1 indicated in Table 6 and depicted in Fig. 1.

Whitney U test versus Student's t test Regarding the use of the Whitney *U* test over the Student's *t* test Zaal Kikvidze et al. [87] evaluated the performance of *t* tests, Mann–Whitney *U* tests, and randomization methods on simulated ecological data with skewed distributions, unequal variances, and unbalanced sample sizes; they found that the Mann–Whitney test failed under unequal variances in large samples, while the *t* test for unequal variances lost power when smaller samples had lower variability.

Datasets

To understand how to properly use the four tests described in this study, we report the results of these tests first on artificial data and then on real-world medical datasets. In this section, we briefly describe the five datasets derived from electronic medical records that we analyze later in this study.

We decided to use data derived from electronic health records (EHRs) in this study for several reasons [88]. First, EHRs are relatively easy to access, especially through open datasets made available for research. Second, they contain clinical information collected for medical purposes, which increases the real-world relevance of the data. Third, EHRs are inherently complex, including heterogeneous variables, missing data, and non-standardized formats. This complexity fit perfectly in the aim of modelling and explaining; finally, working with open EHR data ensures transparency and reproducibility, which are essential for scientific validation.

Original studies on the datasets considered The purpose of this section is to introduce the datasets used to compare the four statistical tests under analysis. These datasets were carefully chosen to demonstrate how each test can be applied effectively in different scenarios, highlighting their strengths and suitability for various types of data.

Yangyan Ma and coauthors [89] analyzed clinical and pathological characteristics, MYCN gene status, surgical methods, and prognosis in neuroblastoma patients from Eastern China. MYCN amplification was associated with significantly lower overall survival, and gross total resection improved survival in advanced-stage cases.

Bulent Gucyetmez et al. [90, 91] analyzed inflammatory markers in over one thousand intensive care patients to evaluate whether C-reactive protein levels and blood count parameters can differentiate sepsis from non-sepsis systemic inflammatory response syndrome. They found that high C-reactive protein levels combined with low lymphocyte and platelet counts significantly increased the likelihood of sepsis, while other common markers such as white blood cell count and neutrophil count were not reliable discriminators.

Bhautesh Dinesh Jani and coauthors [92] investigated the impact of depression on clinical outcomes in heart failure patients from a community cohort. Depression was

associated with increased risks of hospitalization and death, and its inclusion in predictive models significantly improved risk stratification.

Rosa Requena-Morales et al. [93] examined out-of-hospital cardiac arrest mortality in Alicante, Spain, identifying male gender, asystole, longer emergency response time, and cardiac arrest at home as significant risk factors. The findings highlight the need for enhanced CPR training policies to reduce mortality.

Yuichi Takashi and coauthors [94, 95] examined the relationship between serum osteocalcin levels and body fat percentage in Japanese young adults with childhood-onset type one diabetes. The study found an inverse correlation between osteocalcin concentrations and body fat, which remained significant after adjusting for clinical factors. No associations were observed with adiponectin, testosterone, muscle strength, or insulin dose.

Results

In the previous sections, we described the theoretical and mathematical properties and the assumptions of the four statistical tests. Here we move from theory to practice. In this section, we first report the results obtained the the considered statistical tests on artificial data ("Tests on artificial data" section) and then on medical datasets ("Tests on medical data" section).

Tests on artificial data

Before applying the statistical tests to real-world datasets, we performed several sanity checks on artificial data for each test. This step ensures that the tests behave as

Table 7 One sample Student's *t*-test – sanity check. The symbol *** indicates statistical significance at the 0.005 level. Lower *p*-values indicate stronger evidence of a significant difference between the sample and the known mean. As expected, comparisons involving increasingly different values yield progressively smaller *p*-values

Comparison	Batches	Characteristics	*P*-value
identity	sample1 = [2, 2.1, 1.9, 2.2, 1.8]	identical data to the known mean, no difference, the test should not be significant.	1.000
	known mean1 $\mu_0 = 2$		
mean close to the sample	sample2 = [2, 0.1, 3.9, 1.2, 2.8]	known mean close to the sample data, the test does not detect a significant difference.	0.200
	known mean2 $\mu_0 = 3$		
mean far from the sample	sample2 = [2, 0.1, 3.9, 1.2, 2.8]	known mean distant from the sample data, the test detects a significant difference.	0.010
	known mean3 $\mu_0 = 5$		
mean very far from the sample	sample2 = [2, 0.1, 3.9, 1.2, 2.8]	known mean extremely distant from the sample, the test detects a highly significant difference.	3.732×10^{-5}***
	known mean4 $\mu_0 = 15$		

expected under controlled conditions, helping to verify their proper implementation and interpretation.

In Table 7 we report the results of the tests conducted for the one-sample Student's *t*-test using two fixed samples with the same mean, varying the known population mean μ_0 to test the method's responsiveness. To design the sanity checks, we started with samples characterized by low and then higher variance, and tested them against known population means that were initially identical to the sample mean and then progressively more distant, up to values far outside the observed data range. When the known mean equals the sample mean ($\mu_0 = 2$), we expect no difference and a very high *p*-value. This outcome is confirmed by the result $p = 1.000$. With a slightly different known mean ($\mu_0 = 3$), a moderate *p*-value is expected, still above the threshold. The result $p = 0.200$ confirms this. As the known mean increases further ($\mu_0 = 5$), we expect the *p*-value to decrease, approaching significance. Indeed, we obtain $p = 0.010$, consistent with expectations. Finally, with a distant known mean ($\mu_0 = 15$), we expect a highly significant result. The very low value $p = 3.732 \times 10^{-5}$ confirms this. Overall, the results behave as expected: larger discrepancies between the sample and the known mean lead to progressively smaller *p*-values, validating the test.

In Table 8, we disclose the results of the tests carried out for the independent samples Student's *t*-test by comparing different pairs of samples under controlled scenarios. To design the sanity checks, we generated independent sample pairs with increasing differences in variance, and then progressively altered the group means, from identical

Table 8 Independent samples Student's *t*-test – sanity check. The symbol *** indicates statistical significance at the 0.005 level. Lower *p*-values indicate stronger evidence of a significant difference between the two samples. As expected, comparisons involving samples with increasingly different values yield progressively smaller *p*-values

Comparison	Batches	Characteristics	*P*-value
identity	sample1 = [2, 2.1, 1.9, 2.2, 1.8]	identical data, no difference between groups, the test should not be significant.	1.000
	sample1 = [2, 2.1, 1.9, 2.2, 1.8]		
different variances	sample1 = [2, 2.1, 1.9, 2.2, 1.8]	samples with different variances but similar means, the test does not detect a significant difference.	1.000
	sample2 = [2, 0.1, 3.9, 1.2, 2.8]		
highly different variances	sample3 = [35, 45, 50, 55, 65]	samples with highly different variances, the test does not detect a significant difference.	0.872
	sample4 = [–150, 0, 50, 100, 200]		
distant groups	sample5 = [1, 2, 7, 8, 5]	groups with clearly different means, the test detects a significant difference.	2.358×10^{-3}***
	sample6 = [10, 12, 15, 18, 20]		
very distant groups	sample5 = [1, 2, 7, 8, 5]	groups with very distant means and variances, the test detects a highly significant difference.	9.179×10^{-9}***
	sample7 = [60, 65, 68, 63, 70]		

values to increasingly distant ones, to evaluate the test's sensitivity to between-group differences in central tendency. When the two samples are identical, we expect no difference and thus a p-value of 1.000, which is exactly what we observe in the first row of Table 8. In the second case, although the variances differs, the sample means are similar; we expect the test to remain non-significant. The result ($p = 1.000$) confirms this expectation. With two samples showing large differences in variance but still similar central tendencies, we envision a non-significant result again. The obtained p-value of 0.872 confirms that variance alone does not lead to significance in this context. In the fourth case (sample5 and sample6), the two groups have clearly different means. We expect a statistically significant result. The test confirms this expectation by producing $p = 2.358 \times 10^{-3}$, which is below the threshold of 0.005. Finally, when comparing two groups with both very different means and variances, we expect a highly significant difference. The result is indeed strongly significant ($p = 9.179 \times 10^{-9}$), validating the test's sensitivity. Overall, the observed outcomes match theoretical assumptions: as the difference between group means increases, the p-value decreases accordingly, confirming the correct behavior of the independent samples t-test.

In Table 9, we detail the results of the sanity check experiments for the paired samples Student's t-test by comparing fixed sample pairs under increasingly divergent conditions.

Table 9 Paired samples Student's t-test – sanity check. The symbol *** indicates statistical significance at the 0.005 level. Lower p-values indicate stronger evidence of a significant difference between the two samples. As expected, comparisons involving samples with increasingly different values yield progressively smaller p-values.[α]When the paired samples are identical (in our case, sample1 and sample1), all differences between observations are zero. The paired t-test statistic, which divides the average difference by the standard deviation of the differences divided by the square root of the sample size, involves a division by zero. Since the denominator becomes zero, the test is mathematically undefined

Comparison	Batches	Characteristics	*P*-value
identity	sample1 = [2, 2.1, 1.9, 2.2, 1.8]	identical data, no difference between groups, the test should not make sense.	Undefined[α]
	sample1 = [2, 2.1, 1.9, 2.2, 1.8]		
different variances	sample1 = [2, 2.1, 1.9, 2.2, 1.8]	differences in variance in paired data but no significant difference in means.	1.000
	sample2 = [2, 0.1, 3.9, 1.2, 2.8]		
highly different variances	sample3 = [35, 45, 50, 55, 65]	significant variance differences in paired data, the test does not detect a significant difference.	0.859
	sample4 = [–150, 0, 50, 100, 200]		
distant groups	sample5 = [1, 2, 7, 8, 5]	paired data with clearly different means, the test detects a significant difference.	1.000×10^{-3}***
	sample6 = [10, 12, 15, 18, 20]		
very distant groups	sample5 = [1, 2, 7, 8, 5]	paired groups with highly different means, the test detects a highly significant difference.	3.877×10^{-6}***
	sample7 = [60, 65, 68, 63, 70]		

To design the sanity checks, we created paired samples starting from identical values, then varied the within-pair differences by modifying the variance, and finally introduced progressively larger shifts in the paired means to assess how the test responds to increasing paired discrepancies. In the first case, the same sample is compared with itself. Since all paired differences are exactly zero, the test statistic is mathematically undefined due to division by zero. This *undefined* outcome is what we expected to see. In the second case (sample1 and sample2), we compare paired samples with similar means but different variances. Since the differences between pairs are small, we expect a non-significant result. The test confirms this prospect by generating $p = 1.000$. In the third case, the variances differ substantially, but the means are still close. The expected outcome is a insignificant result, which is confirmed by the value $p = 0.859$, clearly above the 0.005 threshold. In the fourth case, the paired samples differ clearly in their means. We expect the test to detect a significant difference. The test returns $p = 1.000 \times 10^{-3}$, below the threshold of 0.005, as forecast. Finally, when the means are very distant (sample5 and sample7), we expect a strongly significant result. The test returns $p = 3.877 \times 10^{-6}$, confirming the expected behavior. These results show that the paired t-test correctly detects differences in means while being robust to variance shifts in the context of paired data. As the magnitude of the mean difference increases, the p-value decreases accordingly.

In Table 10, we outline the results obtained thorughthe sanity checks for the Mann-Whitney U test, by comparing samples under various controlled conditions to assess the test's robustness and limitations. To design the sanity checks, we constructed pairs of independent samples with varying levels of separation, variance, and rank distribution.

Table 10 Mann-Whitney U test – sanity check. The symbol *** indicates statistical significance at the 0.005 level. Lower p-values indicate stronger evidence of a significant difference between the two samples. As expected, comparisons involving samples with increasingly different values yield progressively smaller p-values

Comparison	Samples	Characteristics	*P*-value
identical samples	sample6 = [10, 12, 15, 18, 20]	identical data, no difference between groups, the test should not be significant.	1.000
	sample6 = [10, 12, 15, 18, 20]		
separate samples	sample8 = [1, 1, 1, 1, 1]	significant separation between groups; the test should detect a significant difference.	0.398
	sample9 = [10, 10, 10, 10, 10]		
unequal sample sizes	sample10 = [12, 13, 15]	unequal sample sizes but independent; the test can be used but might be less reliable.	0.024
	sample11 = [20, 22, 21, 23, 25, 19]		
overlapping samples	sample12 = [4, 5, 7, 8, 9]	overlapping data reduces test power; results may not be significant.	0.343
	sample13 = [6, 7, 8, 9, 10]		
asymmetric samples	sample14 = [1, 1, 2, 2, 3, 3]	asymmetric data may distort results; the test may not be significant due to skewed distributions.	4.847×10^{-3}***
	sample15 = [100, 200, 300, 400, 500, 600]		

We began with identical samples, then introduced cases with complete separation but tied ranks, followed by unequal sample sizes, overlapping data, and highly asymmetric distributions, to explore the test's sensitivity and limitations under diverse non-parametric conditions. When the two samples are identical, we expect the test not to detect any difference. The result $p = 1.000$ confirms this belief, as we forecast. In the second case, although the two groups are perfectly separated in value, they have minimal variance and tied ranks, which may limit the test's sensitivity. Despite the large difference in values, the result is $p = 0.398$, not statistically significant, which deviates from expectations and highlights a potential limitation of the test in small, uniform samples. With unequal sample sizes (sample11 and sample12), instead, we expect the test to remain usable but possibly less stable. The result $p = 0.024$ indicates a moderate difference but does not cross the 0.005 threshold, aligning with a lower power due to the imbalance. In the fourth scenario, the two samples overlap considerably. We expect a non-significant result due to reduced rank separation. The test confirms this prediction with $p = 0.343$. Finally, when comparing highly asymmetric samples (sample15 and sample16), we expect a strong signal of difference. Despite the skewness, the Mann-Whitney test returns a significant result ($p = 4.847 \times 10^{-3}$), just below the significance threshold, confirming its sensitivity even in the presence of distributional asymmetry. Overall, the

Table 11 Chi-squared test – sanity check. The symbol *** indicates statistical significance at the 0.005 level. Lower *p*-values indicate stronger evidence of a significant difference between the two samples. As expected, comparisons involving samples with increasingly different values yield progressively smaller *p*-values

Comparison	Batches	Characteristics	*P*-value
uniform distribution	sample16 = [25, 20, 35, 20]	both distributions are close to uniform; the test should not be significant.	0.427
	sample17 = [25, 25, 25, 25]		
imbalanced distribution	sample18 = [10, 12, 8, 10, 10]	slight deviations from uniformity; the test should not be significant.	0.982
	sample9 = [10, 10, 10, 10, 10]		
small sample variation	sample19 = [5, 6, 4, 5, 5, 5]	minimal variance between groups; the test should not be significant.	0.999
	sample20 = [5, 5, 5, 5, 5, 5]		
strong imbalance	sample21 = [4, 10, 4, 10]	strong differences between groups; the test may detect an imbalance.	0.256
	sample17 = [25, 25, 25, 25]		
highly unbalanced	sample22 = [2, 20, 2, 20, 2]	highly skewed distribution; the test should detect a significant difference.	1.000×10^{-7}***
	sample23 = [100, 100, 100, 100, 100]		
extreme imbalance	sample24 = [1, 1, 1, 1, 26]	one category dominates; the test should detect a significant difference.	2.202×10^{-5}***
	sample25 = [6, 6, 6, 6, 6]		

test behaves mostly as expected, though limitations emerge in cases of extreme separation with no variance or when sample sizes are small and homogeneous.

In the Table 11, we outline the results of the sanity check experiments for the Chi-squared test, comparing categorical data distributions to assess the test's sensitivity to imbalance. To design the sanity checks, we created pairs of categorical distributions with increasing imbalance. We started with uniform and nearly uniform frequency distributions, then introduced small and moderate deviations, and finally constructed highly skewed and extreme distributions to assess the test's responsiveness to categorical association under varying levels of discrepancy. When comparing two uniform distributions, we expect no significant difference. The result $p = 0.427$ confirms this. Similarly, in the case of a slight deviation from uniformity, we expect the test to remain non-significant. The test yields $p = 0.982$, which confirms the expectation. When small variations are introduced between groups, as in the third case, we again expect a non-significant result. The value $p = 0.999$ confirms this lack of evidence against the null hypothesis. In the fourth case, the two groups show a strong difference in frequency values, but with small sample sizes. We expect a possible sign of imbalance, though limited power may affect detection. The result $p = 0.256$ indicates no significant difference, aligning with this uncertainty. In the fifth scenario, the distributions are highly skewed. We expect a significant result due to the extreme imbalance. The test confirms this with $p = 1.000 \times 10^{-7}$. Finally, when one group is dominated by a single value and the other is uniform, we expect the test to clearly detect the difference. The test yields $p = 2.202 \times 10^{-5}$, confirming our expectation. Overall, the Chi-squared test behaves as expected: it does not detect differences under uniformity or minimal variation, while it correctly identifies statistically significant discrepancies when the imbalance becomes substantial.

In the Table 12, we report the results achieved through sanity checks for the Kruskal-Wallis test by comparing three independent samples under various conditions, assessing how rank-based differences affect significance. To design the sanity checks, we generated three-group comparisons with increasing differences in central tendency and variance. We started with identical samples, then varied the spread while keeping similar ranks, and finally introduced progressively more distinct mean and rank patterns to evaluate the test's ability to detect rank-based differences across multiple groups. In the first case, all three samples are identical. We expect no rank difference and a non-significant result, which is confirmed by $p = 1.000$. In the second scenario, the samples differ in variance but not in rank distribution. We expect the test not to detect a difference: the p-value of 0.978 confirms assumption. In the third case, despite highly different variances, the samples have nearly identical means and aligned ranks. The test again yields $p = 1.000$, confirming our expectation. When comparing samples with different means and variances but similar rank distributions, we do not expect significance. The test result of $p = 0.645$ aligns with this prospect. In the fifth case, the groups have clearly different means and values spread far apart. We expect to obseverve statistical significance, and the test returns $p = 0.023$, which confirms this although it is just above the stricter threshold of 0.005. With more extreme group differences, we expect strong significance. The result $p = 1.930 \times 10^{-3}$ confirms this prospect. Surprisingly, the final comparison involves similar distributions without evident shifts in ranks, yet the test yields a highly significant result ($p = 2.490 \times 10^{-6}$). This outcome suggests that even subtle, systematic

Table 12 Kruskal-Wallis test – sanity check. The symbol *** indicates statistical significance at the 0.005 level. Lower *p*-values indicate stronger evidence of a significant difference between the two samples. As expected, comparisons involving samples with increasingly different values yield progressively smaller *p*-values

Comparison	Samples	Characteristics	*P*-value
identity	sample1 = [2, 2.1, 1.9, 2.2, 1.8]	identical data, no difference between groups, the test is not significant.	1.000
	sample1 = [2, 2.1, 1.9, 2.2, 1.8]		
	sample1 = [2, 2.1, 1.9, 2.2, 1.8]		
different variances	sample1 = [2, 2.1, 1.9, 2.2, 1.8]	differences in variance, no significant difference in ranks, the test is not significant.	0.978
	sample2 = [2, 0.1, 3.9, 1.2, 2.8]		
	sample26 = [2, 0.1, 0.9, 4.2, 5.8]		
equal means, highly different variances	sample27 = [51.5, 52, 52.5, 53, 53.5]	equal means but highly different variances, no significant difference, the test is not significant.	1.000
	sample28 = [47.5, 50, 52.5, 55, 57.5]		
	sample29 = [10, 30, 52.5, 75, 95]		
different means and variances	sample27 = [51.5, 52, 52.5, 53, 53.5]	different means and variances, no significant difference in ranks, the test is not significant.	0.645
	sample30 = [42.5, 47.5, 50, 52.5, 57.5]		
	sample31 = [10, 30, 50, 70, 90]		
distant groups	sample5 = [1, 2, 7, 8, 5]	paired data with clearly different means, the test is significant.	0.023
	sample7 = [60, 65, 68, 63, 70]		
	sample32 = [252, 22.1, 91.9, 2.2, 41.8]		
very distant groups	sample5 = [1, 2, 7, 8, 5]	highly distant groups, the test is significant.	1.930×10^{-3}***
	sample7 = [60, 65, 68, 63, 70]		
	sample33 = [252, 722.1, 951.9, 782.2, 941.8]		
similar groups	sample34 = [80, 85, 90, 88, 84, 87, 89, 91, 86, 82]	similar groups with no clear difference, the test is highly significant.	2.490×10^{-6}***
	sample35 = [60, 65, 70, 68, 64, 67, 69, 71, 66, 62]		
	sample36 = [40, 45, 50, 48, 44, 47, 49, 51, 46, 42]		

rank shifts can be detected as significant, possibly due to the larger sample size. Overall, the Kruskal-Wallis test behaves as expected in most cases: it remains robust to variance changes when ranks align, and becomes sensitive as rank distributions diverge, although it may sometimes over-detect differences in structured but similar data.

Artificial data results recap. Overall, the results obtained from the sanity checks confirm the theoretical expectations: each statistical test behaved consistently with its

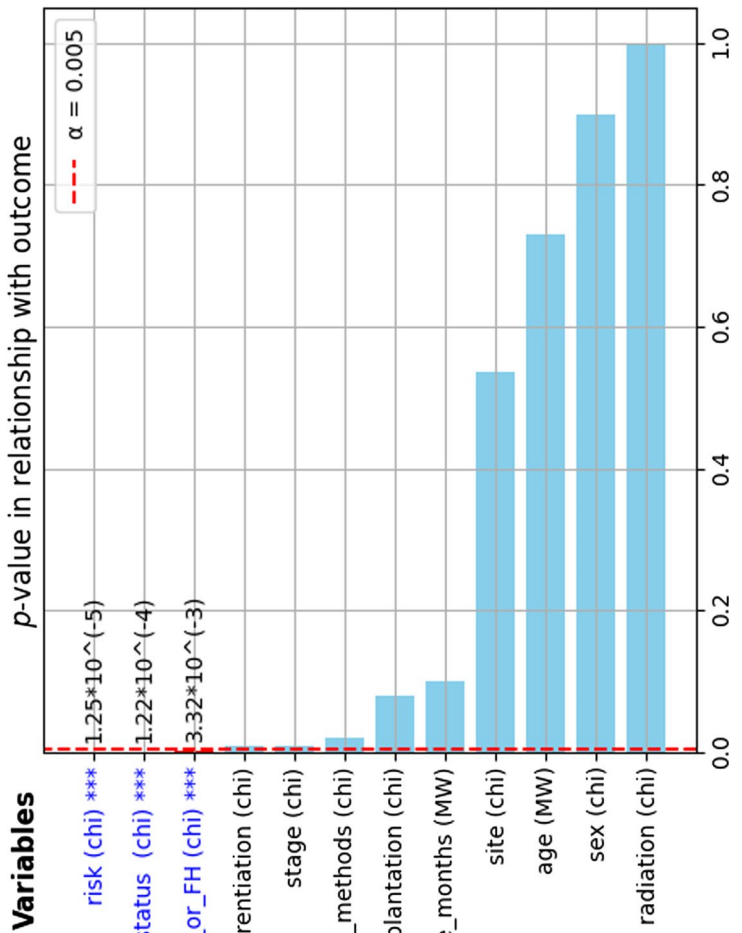

Fig. 3 Barchart of *p*-values – Neuroblastoma dataset. *** indicate significant *p*-values < 0.005 threshold. chi: Chi-squared test. MW: Mann-Whitney *U* test. We reported the meanings of the clinical features in Table 13. More information on this dataset can be found in the original article [89]

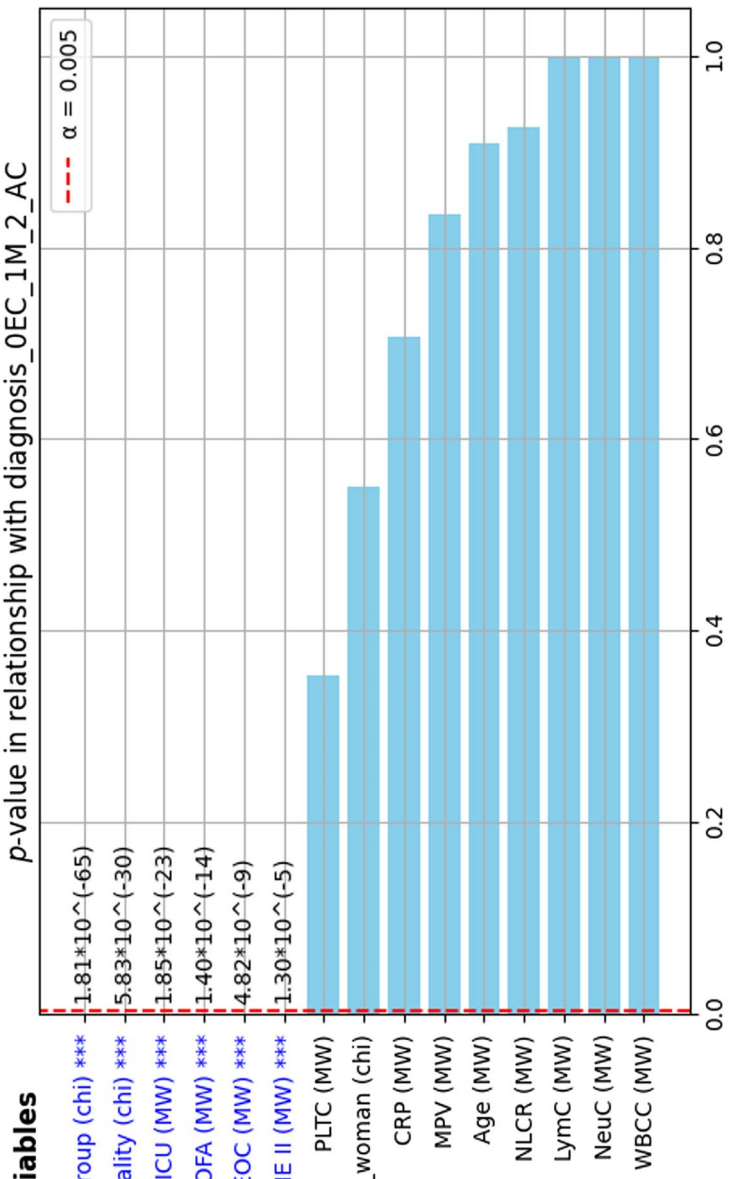

Fig. 4 Barchart of *p*-values – Sepsis and SIRS dataset. *** indicate significant *p*-values < 0.005 threshold. chi: Chi-squared test. MW: Mann-Whitney *U* test. We reported the meanings of the clinical features in Table 14. More information on this dataset can be found in the original article [90]

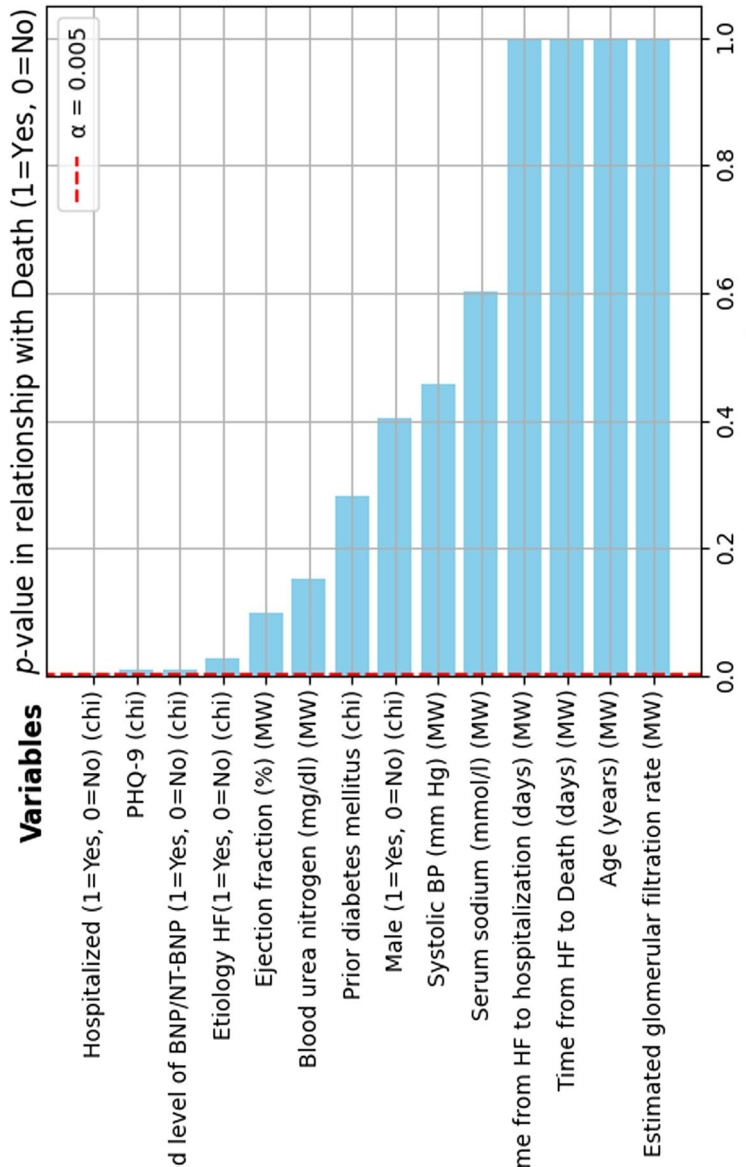

Fig. 5 Barchart of *p*-values – Depression and Heart Failure dataset. *** indicate significant *p*-values < 0.005 threshold. chi: Chi-squared test. MW: Mann-Whitney *U* test. We reported the meanings of the clinical features in Table 15. More information on this dataset can be found in the original article [92]

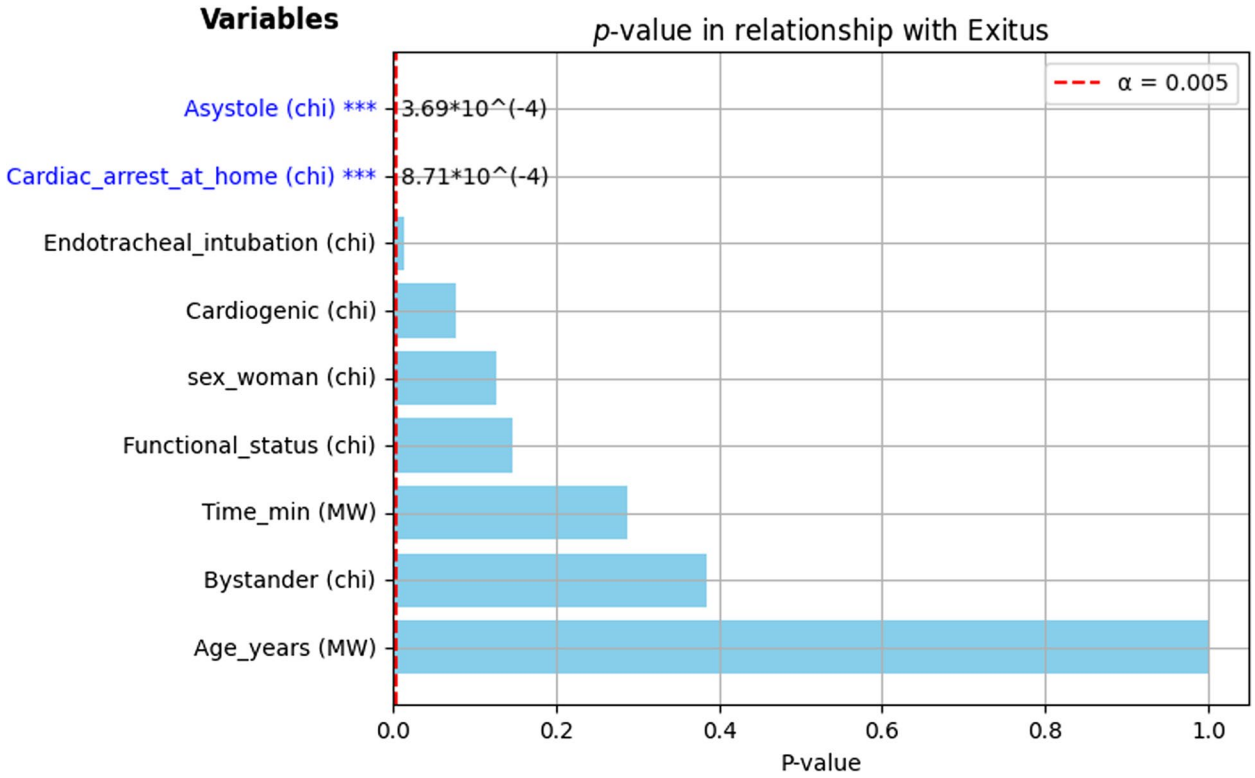

Fig. 6 Barchart of *p*-values – Cardiac Arrest dataset. *** indicate significant *p*-values < 0.005 threshold. chi: Chi-squared test. MW: Mann-Whitney *U* test. We reported the meanings of the clinical features in Table 16. More information on this dataset can be found in the original article [93]

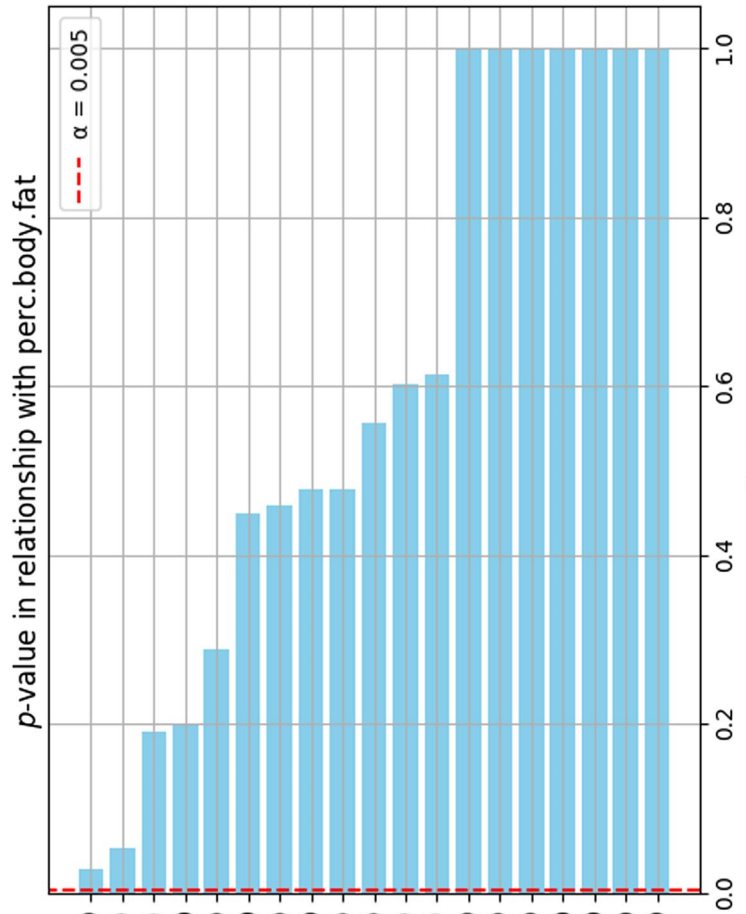

Fig. 7 Barchart of *p*-values – Diabetes Type One dataset. *** indicate significant *p*-values < 0.005 threshold. chi: Chi-squared test. MW: Mann-Whitney *U* test. We reported the meanings of the clinical features in Table 17. More information on this dataset can be found in the original article [94]

underlying assumptions, demonstrating appropriate sensitivity to differences in central tendency, rank distribution, or categorical imbalance, and highlighting specific conditions under which their performance may be limited.

Tests on medical data

The next figures represent barcharts illustrating the p-values obtained from statistical tests (t-test, Chi-squared test, Mann-Whitney U test, Kruskal-Wallis test) conducted on the variables analyzed in relation to the target variable (Figs. 3, 4, 5, 6, 7). Each bar represents the p-value of a variable, ordered from highest to lowest. Bars are colored so that variables with p-values greater than the significance threshold ($p > 0.005$) are shown in sky blue, while those with p-values below the threshold are shown in red, indicating statistical significance.

A dashed line marks the significance threshold to clearly distinguish between significant and non-significant variables. Significant variables are labeled with a star next to their name. Additionally, significant p-values are annotated in scientific notation next to their respective bars.

This visualization allows for a quick identification of variables that significantly impact the outcome of the analysis, making it easier to interpret the determining factors in the statistical evaluation.

Neuroblastoma dataset Figure 3 illustrates the p-values of various variables in relation to the outcome, using Chi-squared tests (chi) and the Mann-Whitney U test (MW). Variables with significant p-values are highlighted in blue, with asterisks denoting their level of significance (*** for p-values <0.005). The red dashed line represents the significance threshold set at $\alpha = 0.005$. The variables "risk", "MYCN status", and "UH or FH" show highly significant relationships with the outcome, with p-values highly below the 0.005 threshold. These findings are consistent with the original study results, which have also identified these variables as significant predictors of the outcome [89]. Additionally, these results are consistent with other published studies: the correlation between the outcome and MYCN status has been verified by Damiano Bartolucci and coauthors [96] and the correlation between outcome and favorable histology (FH) and unfavorable histology (UH) has been verified by Atsuko Nakazawa et al. [97] The remaining variables did not show significance under the stringent $\alpha = 0.005$ threshold: this stricter threshold, compared to the commonly used $\alpha = 0.05$, reduces the likelihood of type one errors, leading to fewer significant results. In the original study of this article [89], the normality threshold has not been specified, and we assume it to be $\alpha = 0.05$. The variables "risk", "MYCN status", and "UH or FH" are highly significant, in agreement with the original publication [89] and confirming prior findings [89, 96, 97].

Sepsis and SIRS dataset Figure 4 depicts the p-values of various clinical and demographic variables in relation to the diagnosis of septic versus non-septic SIRS, using Chi-squared tests (chi) for categorical variables and the Mann-Whitney U test (MW) for continuous or ordinal ones. Variables with significant p-values are highlighted in blue, with asterisks denoting their significance level (*** for p-values < 0.005). The red dashed line indicates the significance threshold set at $\alpha = 0.005$. The variables "Group", "Mortality",

"LOS-ICU", "SOFA", "EOC", and "APACHE II" show strong associations with the diagnosis, with p-values far below the 0.005 threshold. These findings confirm the conclusions of the original study [90], which identified these variables as key markers in discriminating septic from non-septic patients. For example, the correlation between diagnosis and mortality has also been identified by Michael Bauer et al. [98]. The remaining variables, including PLTC, CRP, sex, and various blood cell counts, did not reach statistical significance under the strict $\alpha = 0.005$ threshold. This threshold, which is more conservative than the commonly used $\alpha = 0.05$, reduces false positives but may overlook relevant predictors. Based on the original study [90], variables such as CRP, LymC, and PLTC were nonetheless considered important in multivariate combinations. No normality assumption was stated in the original paper; we assume $\alpha = 0.05$ as the general threshold unless otherwise specified. The variables resulting more relevant through the biostatistics tests are known to be associated with neuroblastoma in the scientific literature.

Depression and Heart Failure dataset Figure 5 shows the p-values of various clinical and demographic variables in relation to two-year mortality in patients diagnosed with heart failure, using Chi-squared tests (chi) for categorical variables and the Mann-Whitney U test (MW) for continuous or ordinal ones. Variables with significant p-values are shown in blue, with asterisks indicating significance levels (*** for p-values <0.005). The red dashed line represents the threshold for significance, set at $\alpha = 0.005$. The variables "Hospitalized", "PHQ-9", "BNP/NT-BNP levels", and "Etiology HF" are associated with highly significant p-values, confirming the original study's findings that depression, hospitalization history, and cardiac biomarkers are strong predictors of mortality [92]. The variables "Hospitalized" ($p = 0.0064$), "PHQ-9" ($p = 0.0118$), "BNP/NT-BNP levels" ($p = 0.0119$), and "Etiology HF" ($p = 0.0302$) did not reach statistical significance under the stricter threshold of $\alpha = 0.005$, but their p-values were close to this cutoff. These results remain directionally consistent with the original study's findings [92], which identified depression, hospitalization history, and cardiac biomarkers as relevant predictors of mortality in heart failure. The correlation between death and Patient Health Questionnaire-9 score (PHQ-9) has also been detected by Scott R. Beach et al. [99]. Other variables such as age, estimated glomerular filtration rate, and time from diagnosis to death or hospitalization were not significant at this threshold. As in the previous dataset, the stricter $\alpha = 0.005$ threshold reduces the probability of false positives, potentially excluding variables that could contribute meaningfully in multivariate settings. The original analysis showed that PHQ-9 in particular improves model performance when included, emphasizing the importance of considering weakly associated variables in broader prediction frameworks [92]. The dataset original study does not explicitly state a normality assumption, so we infer a default threshold of $\alpha = 0.05$.

Cardiac Arrest dataset Figure 6 represents the p-values of various clinical and demographic variables in relation to pre-hospital mortality ("Exitus" variable) following out-of-hospital cardiac arrest, using Chi-squared tests (chi) for categorical variables and the Mann-Whitney U test (MW) for continuous variables. Significant variables are highlighted in blue, and asterisks indicate the level of significance (*** for p-values <0.005). The red dashed line marks the significance threshold, set at $\alpha = 0.005$. The variables

"Asystole" and "Cardiac arrest at home" display p-values well below the threshold, indicating a strong statistical association with the mortality outcome. These findings match those reported in the original study [93], where both conditions were shown to significantly increase the odds of death before hospital arrival. The correlation between "Exitus" and "Asystole" has also been verified by Junki Ishii et al. [100] Other variables like age, sex, time to EMS, and bystander CPR did not reach the stringent threshold, despite some being considered relevant in multivariate analysis. The use of a conservative $\alpha = 0.005$ level limits type I errors but may overlook weaker predictors. The original dataset study [93] also discussed functional status and emergency response delay as important but multifactorial contributors. Normality was not assumed or tested explicitly, so we presume a general threshold of $\alpha = 0.05$.

Diabetes Type One dataset Figure 7 represents the p-values of various biological and clinical variables in relation to body fat percentage in patients with type one diabetes, tested using Chi-squared tests (chi) for categorical variables and the Mann-Whitney U test (MW) for continuous ones. Significant p-values are shown in blue, and levels of significance are indicated by asterisks (*** for p-values <0.005). The red dashed line marks the threshold for significance at $\alpha = 0.005$. The variables "body mass index", "sex", and "HbA1c" exhibit strong associations with the body fat outcome, with p-values well below the threshold. These results are consistent with the original study [94], which emphasized the role of glycemic control and anthropometric factors in fat accumulation. The correlation between percent body fat and HbA1c has also been confirmed by a study by Julie Bower et al. [101]. None of the tested variables, including "body mass index", "sex", and "HbA1c", reached statistical significance at the predefined threshold of $\alpha = 0.005$. Nonetheless, body mass index ($p = 0.0306$) and sex ($p = 0.055$) showed p-values closer to the cutoff and may still be of interest. These findings are only partially consistent with the original study [94], which emphasized the influence of glycemic control and anthropometric factors on fat accumulation. While the association between percent body fat and HbA1c has been reported elsewhere [101], it was not confirmed in our analysis. Other variables, including ucOC and OC, did not reach significance in this univariate analysis, but were found to have an inverse relationship with body fat in the original multivariate regression. Additional variables such as adiponectin, testosterone, and insulin dosage also failed to show significance under the $\alpha = 0.005$ criterion. As with other datasets, this stricter threshold reduces type I error risk, though it may exclude weak but meaningful predictors. The original dataset study [94] does not specify any test for normality, so we assume a conventional $\alpha = 0.05$ baseline.

EHRs results recap scrivere qua riassunto Using a significance threshold of $\alpha = 0.005$, only a subset of variables reached statistical significance in univariate tests. However, several others, such as "PHQ-9" in heart failure and "Body Mass Index" in type one diabetes, showed p-values close to the cutoff and may still hold predictive value in multivariate models. Overall, the results align with the original studies: key predictors such as "MYCN status" in neuroblastoma and "SOFA" in sepsis were confirmed. Some expected

associations did not reach significance, likely due to the conservative threshold and the univariate nature of the analysis. In absence of normality assumptions in the original sources, non-parametric methods were used, and comparisons with the conventional $\alpha = 0.05$ suggest that certain borderline variables could be reconsidered in broader modeling frameworks. Consistencies with external studies support the clinical relevance of several findings, even when p-values were not below the threshold.

Discussion and conclusions

A manual on univariate statistical tests In an era where machine learning and artificial intelligence are becoming increasingly prevalent, biostatistics can serve as a useful and simpler tool for assessing the validity of scientific results obtained through computational analyses. Biostatistics offers a range of statistical tests that are low in computational cost and can be executed quickly, providing researchers with insights into potential strong relationships between two samples. However, these statistical tests are often chosen incorrectly, applied out of context, or used in a misleading manner, particularly in biomedical informatics. In this study, we try to alleviate this problem by analyzing the four most common univariate statistical tests that we have employed and observed in our scientific activities: Student's t-test, Chi-squared test, Kruskal-Wallis test, and Mann-Whitney U test. Our goal is to bring some order to this topic and provide a general guide on when to choose one test over another.

We first introduce these four biostatistics tests by explaining their mathematical properties, assumptions, produced statistics, and the meanings of their p-values. Then we provide some guidelines on when to choose which test, based on the datatypes of the samples considered. Afterwards, moving from theory to practice we applied these four tests to artificial data and to real-world medical data derived from electronic medical

Table 13 Description of the variables used in the Neuroblastoma dataset. The target variable we consider is "outcome", used to assess survival

Variable	Description and possible values	Type
Age	Any positive integer	Numerical
Sex	0 = Female, 1 = Male	Categorical
Site	0, 1, 2,... (codes for tumor primary sites)	Categorical
Stage	1, 2, 3, 4	Ordinal
Risk	0 = Low risk, 1 = High risk	Categorical
time months	Any positive integer	Numerical
autologous stem cell transplantation	0 = No, 1 = Yes	Categorical
radiation	0 = No, 1 = Yes	Categorical
degree of differentiation	0 = Well differentiated, 1 = Moderately differentiated, 2 = Poorly differentiated	Ordinal
UH or FH	0 = Absent, 1 = Present	Categorical
MYCN status	0 = Normal, 1 = Amplified	Categorical
surgical methods	0 = No surgery, 1 = Surgery performed	Categorical
outcome	0 = Not survived, 1 = Survived	Categorical

Table 14 Description of the variables used in the Sepsis & SIRS dataset. The target variable we consider is "diagnosis_0EC_1M_2AC", used to assess the diagnosis in the patients

Variable	Description and possible values	Type
Age	Patient's age in years (positive integers)	Numerical
sex_woman	0 = Male, 1 = Female	Categorical
diagnosis 0EC 1M 2AC	0 = Elective surgery, 1 = Medical disease, 2 = Emergency surgery	Categorical
APACHE II	Acute Physiology and Chronic Health Evaluation score at ICU admission	Numerical
SOFA	Sequential Organ Failure Assessment score at ICU admission	Numerical
CRP	C-Reactive Protein measured in mg/dL at ICU admission	Numerical
WBCC	White Blood Cell Count measured in 10^3 cells/μL at ICU admission	Numerical
NeuC	Neutrophil Count measured in 10^3 cells/μL at ICU admission	Numerical
LymC	Lymphocyte Count measured in 10^3 cells/μL at ICU admission	Numerical
EOC	Eosinophil Count measured in cells/μL at ICU admission	Numerical
NLCR	Neutrophil-to-Lymphocyte Count Ratio at ICU admission	Numerical
PLTC	Platelet Count measured in 10^3 cells/μL at ICU admission	Numerical
MPV	Mean Platelet Volume measured in femtoliters (fL) at ICU admission	Numerical
Group	0 = Non-sepsis SIRS, 1 = Sepsis	Categorical
LOS-ICU	Length of ICU stay in days (positive integers)	Numerical
Mortality	0 = Alive at ICU discharge, 1 = Deceased during ICU stay	Categorical

Table 15 Description of the variables used in the Depression and Heart Failure dataset. The target variable we consider is "Death", used to assess death within two years

Variable	Description and possible values	Type
Age (years)	Age in years	Numerical
Male	Patient gender (1 = Male, 0 = Female)	Categorical
PHQ-9	Patient Health Questionnaire score	Ordinal
Systolic BP (mm Hg)	Systolic blood pressure in mm Hg	Numerical
Estimated glomerular filtration rate	Estimated glomerular filtration rate	Numerical
Ejection fraction (%)	Heart ejection fraction (percentage)	Numerical
Serum sodium (mmol/l)	Blood sodium concentration in mmol/l	Numerical
Blood urea nitrogen (mg/dl)	Blood urea nitrogen in mg/dl	Numerical
Etiology HF	Etiology of heart failure (1 = Yes, 0 = No)	Categorical
Prior diabetes mellitus	Presence of pre-existing diabetes mellitus	Categorical
Elevated level of BNP/NT-BNP	Elevated BNP or NT-BNP levels (1 = Yes, 0 = No)	Categorical
Time from HF to Death (days)	Time from heart failure to death (in days)	Numerical
Death	Death within two years (1 = Yes, 0 = No)	Categorical
Time from HF to hospitalization (days)	Time from heart failure to hospitalization (in days)	Numerical
Hospitalized	Hospitalization within two years (1 = Yes, 0 = No)	Categorical

records. This way, we displayed the behavior of the analyzed four tests on a practical application.

This study has highlighted that the choice of statistical test is not arbitrary but instead highly dependent on several key factors. Among these, the type of data—whether numerical, categorical, or ordinal—plays a crucial role. Particular attention should be given to binary variables, for which the decision to treat them as categorical or ordinal depends on both their nature and the context of analysis. Additionally, the normality of numerical data and the sample size significantly influence which test is most

Table 16 Description of the variables used in the Cardiac Arrest dataset. The target variable we consider is "Exitus", used to assess survival before hospital arrival

Variable	Description and possible values	Type
Exitus	Outcome of cardiac arrest (0 = Deceased, 1 = Survived)	Categorical
sex woman	Patient's sex (0 = Male, 1 = Female)	Categorical
Age years	Patient's age in years (positive integers)	Numerical
Endotracheal intubation	Endotracheal intubation performed (0 = No, 1 = Yes)	Categorical
Functional_status	Pre-arrest functional status (scale from 0 to 3)	Ordinal
Asystole	Presence of asystole (0 = No, 1 = Yes)	Categorical
Cardiac arrest at home	Location of cardiac arrest (0 = Other, 1 = Home)	Categorical
Bystander	Bystander CPR provided (0 = No, 1 = Yes)	Categorical
Time min	Time from arrest to EMS arrival (in minutes)	Numerical
Cardiogenic	Cardiac etiology of arrest (0 = No, 1 = Yes)	Categorical

Table 17 Description of the variables used in the Diabete Type One dataset. The target variable we consider is "perc.body.fat", used to assess the body fat percentage

Variable	Description and possible values	Type
age	Patient's age in years	Numerical
duration of diabetes	Duration of diabetes in years	Numerical
body mass index	Body Mass Index (BMI)	Numerical
TDD	Total daily dose of insulin	Numerical
basal	Basal insulin dose	Numerical
bolus	Rapid-acting (bolus) insulin dose	Numerical
HbA1c	Glycated hemoglobin	Numerical
eGFR	Estimated glomerular filtration rate	Numerical
perc body fat	Body fat percentage	Numerical
adiponectin	Serum adiponectin level	Numerical
free testosterone	Serum free testosterone level	Numerical
SMI	Skeletal muscle mass index	Numerical
grip strength	Hand grip strength	Numerical
knee extension strength	Knee extension strength	Numerical
gait speed	Walking speed	Numerical
ucOC	Serum undercarboxylated osteocalcin level	Numerical
OC	Serum osteocalcin level	Numerical
weight kg	Body weight in kilograms	Numerical
insulin regimen binary	Insulin regimen (0 = No insulin, 1 = Insulin)	Categorical
sex 0man 1woman	Sex (0 = Male, 1 = Female)	Categorical

appropriate. We believe our study can have a strong impact on the scientific community: potentially, anyone about to use a biostatistics univariate test will be able to take advantage of our study guide to make the proper selection. This idoneous choice, in turn, can help produce more reliable and robust scientific results, increasing the quality of scientific findings in any field. In theory, our study can serve thousands of scientific researchers worldwide.

We invite anyone performing a scientific study involving univariate statistical tests to take into consideration this present guide, and to use it to carefully choose the

most suitable test for their case. The flowchart listed before (Figs. 1 and 2) should guide any researcher before making a decision on which univariate statistical test to employ in a specific scenario-

Limitations Several limitations of the study must be acknowledged. First, the analysis was restricted to data derived exclusively from electronic health records, thereby excluding other potentially informative sources. Second, the study focused on only five datasets, which constitutes a relatively small sample and may limit the robustness and generalizability of the findings.

Our choice of using 50 as a threshold for the normality testing can also be questioned. We selected this number based on a traditional choice carried out in the biostatistics community [85, 86], but we acknowledge that other values can be utilized for this purpose as well.

The scope of the study was intentionally narrowed to just four well-known univariate tests: the Student's t-test, the Mann-Whitney U test, the Chi-squared test, and the Kruskal-Wallis test. This focus excluded other valid approaches that might be suitable for specific data types or research questions. Furthermore, we did not comnpare the results of our tests on artificial data and on medical data with any ground truth, to quantify how much they were right. We did not have this possibility because a real ground truth for these data does not exist. For the medical data, we found evidence about the associations between clinical features in the scientific literature, but we could not state how correct the tests' p-values were.

Future developments Future research could aim to generalize these results beyond the medical domain and apply them to datasets not derived from electronic health records. Expanding the range of tests and software tools considered could also provide a more comprehensive evaluation framework and offer deeper insights into the practical implications of test selection in applied research. The present guide is focused on the univariate tests; we envision a future new study on multivariate tests, such as the Hotelling's t-squared test [102].

Abbreviations

ALT	Alanine aminotransferase
ANOVA	Analysis of variance
BMI	Body mass index
BRCA	Breast cancer gene
CPR	Cardiopulmonary resuscitation
COVID-19	Coronavirus disease 2019
DNA	Deoxyribonucleic acid
EHRs	Electronic health records
FH	Familial hypercholesterolemia
FIT	Fecal immunochemical test
GCA	Gastric cardia adenocarcinoma
GPL-3	GNU General Public License version 3
HER2	Human epidermal growth factor receptor 2
HIE	Health information exchange
HPV	Human papillomavirus
IDC	Invasive ductal carcinoma
IMPC	Invasive micropapillary carcinoma

KW	Kruskal-Wallis test
LDL-C	Low-density lipoprotein cholesterol
MS	Multiple sclerosis
MYCN	MYCN Proto-Oncogene, BHLH Transcription Factor
MW	Mann-Whitney *U* test
PACU	Post-anesthesia care unit
PHQ-9	Patient health questionnaire-9
PROM	Patient-reported outcome measurement
ROC	Receiver operating characteristic curve
SL	Serum levels
SSIs	Surgical site infections
TCGA	The Cancer Genome Atlas
TMAO	Trimethylamine n-oxide
WM	Weight management

Acknowledgements
The authors acknowldedge the use of Ecosia AI Chat for English proof-reading of the text of this article.

Software code availability
Our software code is publicly available under the GPL-3.0 license on GitHub at https://github.com/AndreaSichenze/Biostatistics_tests and on Zenodo at https://doi.org/10.5281/zenodo.15544014.

Authors' contributions
D.C. conceived and supervised the study, contributed to the writing of the manuscript, and reviewed the manuscript. A.S. performed the tests, wrote the software code, designed the experiments, wrote the literature review, wrote several parts of the manuscript, and reviewed the manuscript. G.J. supervised the study, provided feedback on the article's contents, and reviewed the manuscript.

Funding
The work of D.C. is partially funded by the Italian Ministero Italiano delle Imprese e del Made in Italy under the Digital Intervention in Psychiatric and Psychologist Services (DIPPS) programme and is partially supported by Ministero dell'Università e della Ricerca of Italy under the "Dipartimenti di Eccellenza 2023-2027" ReGAInS grant assigned to Dipartimento di Informatica Sistemistica e Comunicazione at Università di Milano-Bicocca. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data availability
Data availability The EHRs datasets employed in this study are publicly available in the supplementary ma- terials of the original dataset publications [83, 84, 86–88] under the CC BY 4.0 license at the following URLs:
– Sepsis and SIRS [84] https://figshare.com/articles/dataset/_C_Reactive_Protein_and_Hemogram_Parameters_for_the_Non_Sepsis_Systemic_Inflammatory_Response_Syndrome_and_Sepsis_What_Do_They_Mean_/1644426?file=2637248
– Depression and Heart Failure [86] https://figshare.com/articles/dataset/Comorbid_Depression_and_Heart_Failure_A_Community_Cohort_Study/3916224?file=6130425
– Cardiac Arrest [87] https://figshare.com/articles/dataset/Mortality_after_out-of-hospital_cardiac_arrest_in_a_Spanish_Region/4876247?file=8166893
– Neuroblastoma [83] https://doi.org/10.7717/peerj.5665/supp-5
– Diabetes Type One [88] https://figshare.com/articles/dataset/Circulating_osteocalcin_as_a_bone-derived_hormone_is_inversely_correlated_with_body_fat_in_patients_with_type_1_diabetes/8079389?file=15057092.

Declarations

Ethics approval and consent to participate
Permission to collect and analyze the data of patients' involved in this study has been obtained by the original datasets curators from ethical committees of their hospitals, as stated in the [89, 90, 92–94] original articles.

Competing interests
The authors declare no competing interests.

Received: 3 June 2025 Accepted: 20 July 2025
Published online: 20 August 2025

References

1. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380(14):1347–58. https://doi.org/10.1056/nejmra1814259.
2. Alves R, Pasquier C, Pasquier N. The pervasiveness of machine learning in omics science. In: ECML PKDD 2014 – International Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases Tutorial T3. 2014. https://hal.science/hal-01330594v1. Accessed 1st July 2025.

3. Sambasivan N, Kapania S, Highfill H, Akrong D, Paritosh P, Aroyo LM. "Everyone wants to do the model work, not the data work": data cascades in High-Stakes AI. In: Proceedings of CHI '21 – the 2021 CHI Conference on Human Factors in Computing Systems. ACM; 2021. pp. 1–15. https://doi.org/10.1145/3411764.3445518.
4. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intel. 2019;1(5):206–15. https://doi.org/10.1038/s42256-019-0048-x.
5. Gosset WSS. The probable error of a mean. Biometrika. 1908:1–25. https://doi.org/10.2307/2331554.
6. DATAtab Team. t-Test. 2025. https://datatab.net/tutorial/t-test. URL visited on 2nd May 2025. Accessed 1st July 2025.
7. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat. 1947:50–60. https://www.jstor.org/stable/2236101. Accessed 1st July 2025.
8. DATAtab Team. Mann-Whitney U-Test. 2025. https://datatab.net/tutorial/mann-whitney-u-test. URL visited on 2nd May 2025. Accessed 1st July 2025.
9. Cochran WG. The χ^2 Test of Goodness of Fit. Ann Math Stat. 1952;23:315–345. https://api.semanticscholar.org/CorpusID:121913312. Accessed 1st July 2025.
10. DATAtab Team. Chi-Square test. 2025. https://datatab.net/tutorial/chi-square-test. URL visited on 2nd May 2025. Accessed 1st July 2025.
11. Kruskal WH, Wallis WA. Use of Ranks in One-Criterion Variance Analysis. J Am Stat Assoc. 1952;47:583–621. https://api.semanticscholar.org/CorpusID:51902974. Accessed 1st July 2025.
12. DATAtab Team. Kruskal-Wallis-Test. 2025. https://datatab.net/tutorial/kruskal-wallis-test. URL visited on 2nd May 2025. Accessed 1st July 2025.
13. Moynihan D, Monaco S, Ting TW, Narasimhalu K, Hsieh J, Kam S, et al. Author Correction: Analysis and visualisation of electronic health records data to identify undiagnosed patients with rare genetic diseases. Sci Rep. 2024;14(1):10084. https://doi.org/10.1038/s41598-024-55424-8.
14. Patton H, Burchette R, Tovar S, Pio J, Shi J, Nyberg LM. Retrospective analysis of a dedicated care pathway for non-alcoholic fatty liver disease in an integrated US healthcare system demonstrates support of weight management and improved ALT. BMC Gastroenterol. 2020;20(1):362. https://doi.org/10.1186/s12876-020-01492-9.
15. Blair CK, Wiggins CL, Nibbe AM, Storlie CB, Prossnitz ER, Royce M, et al. Obesity and survival among a cohort of breast cancer patients is partially mediated by tumor characteristics. NPJ Breast Cancer. 2019;5(1):33. https://doi.org/10.1038/s41523-019-0128-4.
16. Nelson CA, Bove R, Butte AJ, Baranzini SE. Embedding electronic health records onto a knowledge network recognizes prodromal features of multiple sclerosis and predicts diagnosis. J Am Med Inform Assoc. 2021;29(3):424–34. https://doi.org/10.1093/jamia/ocab270.
17. David CH, Quessard A, Mastroianni C, Hekimian G, Amour J, Leprince P, et al. Mechanical circulatory support with the Impella 5.0 and the Impella Left Direct pumps for postcardiotomy cardiogenic shock at La Pitié-Salpêtrière Hospital. Eur J Cardiothorac Surg. 2019;57(1):183–188. https://doi.org/10.1093/ejcts/ezz179.
18. Gao J, Yan KT, Wang JX, Dou J, Wang J, Ren M, et al. Gut microbial taxa as potential predictive biomarkers for acute coronary syndrome and post-STEMI cardiovascular events. Sci Rep. 2020;10(1):2639. https://doi.org/10.1038/s41598-020-59235-5.
19. Issaka RB, Singh MH, Oshima SM, Laleau VJ, Rachocki CD, Chen EH, et al. Inadequate Utilization of Diagnostic Colonoscopy Following Abnormal FIT Results in an Integrated Safety-Net System. Am J Gastroenterol. 2016;112(2):375–82. https://doi.org/10.1038/ajg.2016.555.
20. Storm AC, Ryou M, Thompson CC. Multicenter implementation of a new electronic medical record system leads to longer procedure times and poor staff satisfaction. Clin Endosc. 2019;52(1):87–9. https://doi.org/10.5946/ce.2018.080.
21. Zhou JZ, Liu X, Ye GJ. The impact of workplace bullying on depression among clinical nurses in China: a comparative analysis. Medicine. 2025;104(2). https://doi.org/10.1097/md.0000000000041246.
22. Wang G, Liu L, Zhao Y, Lin Y, Er L. Comparative genomic analysis unveiling the mutational landscape associated with premalignant lesions and early-stage gastric cardia cancer. Medicine. 2025;104(2). https://doi.org/10.1097/md.0000000000040332.
23. Do TM, Nguyen QHN, Le NHD, Nguyen HD, Phung AHT, Tran TS, et al. Association between dietary factors and breast cancer risk: a matched case-control study in Vietnam. BMC Cancer. 2024;24(1):1224. https://doi.org/10.1186/s12885-024-12918-y.
24. Lapin BR, Honomichl RD, Thompson NR, Rose S, Sugano D, Udeh B, et al. Association Between Patient Experience With Patient-Reported Outcome Measurements and Overall Satisfaction With Care in Neurology. Value Health. 2019;22(5):555–63. https://doi.org/10.1016/j.jval.2019.02.007.
25. Lugo-Perez S, Azpiri-Lopez JR, Galarza-Delgado DA, Colunga-Pedraza IJ, Cardenas-De La Garza JA, Arvizu-Rivera RI, et al. Cardiac remodeling in rheumatoid arthritis: assessing the influence of anti-CCP and rheumatoid factor. Eur Heart J. 2024;45(Supplement_1):ehae666.2729. https://doi.org/10.1093/eurheartj/ehae666.2729.
26. Steinbruck I, Ebigbo A, Kuellmer A, Schmidt A, Kouladouros K, Brand M, et al. Cold Versus Hot Snare Endoscopic Resection of Large Nonpedunculated Colorectal Polyps: randomized Controlled German CHRONICLE Trial. Gastroenterology. 2024;167(4):764–77. https://doi.org/10.1053/j.gastro.2024.05.013.
27. Qian Y, Yuan L, Zhang X. Comparative study on blood pressure and metabolic improvements in hypertensive patients using copper bianstone scraping. Medicine. 2025;104(2). https://doi.org/10.1097/md.0000000000041133.
28. Meng J, Zhu Y, Li Y, Sun T, Zhang F, Qin S, et al. Incidence and risk factors for surgical site infection following elective foot and ankle surgery: a retrospective study. J Orthop Surg Res. 2020;15(1):449. https://doi.org/10.1186/s13018-020-01972-4.
29. Lapin B, Udeh B, Bautista JF, Katzan IL. Patient experience with patient-reported outcome measures in neurologic practice. Neurology. 2018;91(12):e1135–51. https://doi.org/10.1212/wnl.0000000000006198.
30. Esaki M, Suzuki S, Hayashi Y, Yokoyama A, Abe S, Hosokawa T, et al. Propensity score-matching analysis to compare clinical outcomes of endoscopic submucosal dissection for early gastric cancer in the postoperative and non-operative stomachs. BMC Gastroenterol. 2018;18(1):125. https://doi.org/10.1186/s12876-018-0855-2.

31. Zelviene A, Bogusevicius A. Reliability and validity of the Champion's Health Belief Model Scale among Lithuanian women. Cancer Nurs. 2007;30(3):E20-8. https://doi.org/10.1097/01.ncc.0000270711.72413.a6.
32. Cerulo L, Pagnotta SM. massiveGST: a Mann-Whitney-Wilcoxon gene-set test tool that gives meaning to gene-set enrichment analysis. Entropy. 2022;24(5):739. https://doi.org/10.3390/e24050739.
33. Zhu S, Tong Y, Chen W, Chen X, Shen K. Association of obesity and luminal subtypes in prognosis and adjuvant endocrine treatment effectiveness prediction in Chinese breast cancer patients. 2021. https://doi.org/10.3389/fonc.2022.862224.
34. Zhao D, Wang X, Beeraka NM, Zhou R, Zhang H, Liu Y, et al. High body mass index was associated with human epidermal growth factor receptor 2-positivity, histological grade and disease progression differently by age. World J Oncol. 2023;14(1):75–83. https://doi.org/10.14740/wjon1543.
35. Leone JP, Leone J, Hassett MJ, Freedman RA, Avila J, Vallejo CT, et al. Incidence, treatment patterns, and mortality for patients with breast cancer during the first year of the COVID-19 pandemic: a population-based study. Breast Cancer Res Treat. 2024. https://doi.org/10.1007/s10549-024-07562-w.
36. RECOVERY Collaborative Group. Convalescent plasma in patients admitted to hospital with COVID-19 (RECOVERY): a randomised controlled, open-label, platform trial. Lancet. 2021;397(10289):2049–59. https://doi.org/10.1016/S0140-6736(21)00897-7.
37. Jin J, Li B, Cao J, Li T, Zhang J, Cao J, et al. Analysis of clinical features, genomic landscapes and survival outcomes in HER2-low breast cancer. J Transl Med. 2023;21(1):360. https://doi.org/10.1186/s12967-023-04076-9.
38. Magnoni F, Bianchi B, Pagan E, Corso G, Sala I, Bagnardi V, et al. Long-term outcome of invasive pure micropapillary breast cancer compared with invasive mixed micropapillary and invasive ductal breast cancer: a matched retrospective study. Breast Cancer Res Treat. 2024;208(2):333–47. https://doi.org/10.1007/s10549-024-07422-7.
39. Tarı Selçuk K, Avcı D, Yılmaz Dündar G, Mercan Y. Breast cancer screening behaviors in women aged 40 years and over in a semi-urban region in Turkey: relationships with health beliefs. Healthcare. 2020;8(2):171. https://doi.org/10.3390/healthcare8020171.
40. Bhattacharyya O, Rawl SM, Dickinson SL, Haggstrom DA. Comparison of health information exchange data with self-report in measuring cancer screening. BMC Med Res Methodol. 2023;23(1):172. https://doi.org/10.1186/s12874-023-01907-7.
41. Cengiz B, Bahar Z, Canda AE. The effects of patient care results of applied nursing intervention to individuals with stoma according to the Health Belief Model. Cancer Nurs. 2020;43(2):E87–96. https://doi.org/10.1097/ncc.0000000000000678.
42. Wu Y, Fan J, Peissig P, Berg R, Tafti AP, Yin J, et al. Quantifying predictive capability of electronic health records for the most harmful breast cancer. In: Medical Imaging 2018: Image Perception, Observer Performance, and Technology Assessment, vol. 10577. SPIE; 2018. pp. 112–120. https://doi.org/10.1117/12.2293954.
43. Clifton JC, Ende HB, Rathnam C, Freundlich RE, Sandberg WS, Wanderer JP. A mobile post anesthesia care unit order reminder system improves timely order entry. J Med Syst. 2024;48(1):60. https://doi.org/10.1007/s10916-024-02079-7.
44. Karajizadeh M, Nikandish R, Zalpour Z, Roozrokh Arshadi Montazer M, Soleimanijafarbiglo M, Mazaher Y, et al. Identification of the Information Needs of a Nurse-led Rapid Response Team to Design and Develop an Electronic Medical Record System. Health Manag Inf Sci. 2022;9(4):236–242. https://doi.org/10.30476/jhmi.2023.96214.1143.
45. Dexter F, Pinho RH, Pang DSJ. Modeling daily veterinary anesthetist patient care hours and probabilities of exceeding critical thresholds. Am J Vet Res. 2024;85(5):1–10. https://doi.org/10.2460/ajvr.23.09.0196.
46. Katzan IL, Thompson NR, Dunphy C, Urchek J, Lapin B. Neurologic provider views on patient-reported outcomes including depression screening. Neurol Clin Pract. 2018;8(2):86–92. https://doi.org/10.1212/cpj.0000000000000438.
47. Williams DC, Warren RW, Ebeling M, Andrews AL, Teufel RJ II. Physician use of electronic health records: Survey study assessing factors associated with provider reported satisfaction and perceived patient impact. JMIR Med Inform. 2019;7(2):e10949. https://doi.org/10.2196/10949.
48. Jabali AK. Predictors of Anesthesiologists' attitude toward EHRs in Saudi Arabia for clinical practice. Inform Med Unlocked. 2021;23(100555):100555. https://doi.org/10.3389/fdgth.2023.1252227.
49. Fors M, Ballaz S, Ramírez H, Mora FX, Pulgar-Sánchez M, Chamorro K, et al. Sex-dependent performance of the neutrophil-to-lymphocyte, monocyte-to-lymphocyte, platelet-to-lymphocyte and mean platelet volume-to-platelet ratios in discriminating COVID-19 severity. Front Cardiovasc Med. 2022;9:822556. https://doi.org/10.3389/fcvm.2022.822556.
50. Tollinche LE, Shi R, Hannum M, McCormick P, Thorne A, Tan KS, et al. The impact of real-time clinical alerts on the compliance of anesthesia documentation: A retrospective observational study. Comput Methods Prog Biomed. 2020;191(105399):105399. https://doi.org/10.1016/j.cmpb.2020.105399.
51. Shafran-Tikva S, Gabay G, Kagan I. Transformative insights into community-acquired pressure injuries among the elderly: a big data analysis. Healthcare. 2025;13(2). https://doi.org/10.3390/healthcare13020153.
52. Makin TR, Orban de Xivry JJ. Ten common statistical mistakes to watch out for when writing or reviewing a manuscript. eLife. 2019;8:e48175. https://doi.org/10.7554/elife.48175.
53. D'Arrigo G, El Hafeez SA, Mezzatesta S, Abelardo D, Provenzano FP, Vilasi A, et al. Common mistakes in biostatistics. Clin Kidney J. 2024;17(7):sfae197. https://doi.org/10.1093/ckj/sfae197.
54. AbdulRaheem Y. Statistics in medical research: common mistakes. J Taibah Univ Med Sci. 2023;18(6):1197. https://doi.org/10.1016/j.jtumed.2023.04.004.
55. Hanna M. Statistics: common mistakes. How to Write Better Medical Papers. 2019:73–81. https://doi.org/10.1007/978-3-030-02955-5_14.
56. Cowger CD. Correcting misuse is the best defense of statistical tests of significance. Soc Serv Rev. 1987;61(1):170–172. https://www.jstor.org/stable/30011875. Accessed 1st July 2025.
57. Streiner DL. Best (but oft-forgotten) practices: the multiple problems of multiplicity–whether and how to correct for many statistical tests. Am J Clin Nutr. 2015;102(4):721–8. https://doi.org/10.3945/ajcn.115.113548.

58. Schatz P, Jay KA, McComb J, McLaughlin J. Misuse of statistical tests in publications. Arch Clin Neuropsychol. 2005;20(8):1053–9. https://doi.org/10.1016/j.acn.2005.06.006.
59. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, et al. Redefine statistical significance. Nat Hum Behav. 2018;2(1):6–10. https://doi.org/10.1038/s41562-017-0189-z.
60. Khan Academy. Examples of null and alternative hypotheses. 2025. https://www.khanacademy.org/math/ap-statistics/xfb5d8e68:inference-categorical-proportions/idea-significance-tests/v/examples-of-null-and-alternative-hypotheses. URL visited on 2nd May. Accessed 1st July 2025.
61. Levene H. Robust tests for equality of variances. Contributions to probability and statistics. Technical Report, Stanford University, 1960:278–92.
62. Conover WJ. Practical Nonparametric Statistics, vol. 350. John Wiley & Sons; 1999. https://www.wiley.com/en-us/Practical+Nonparametric+Statistics%2C+3rd+Edition-p-9780471160687. Accessed 1st July 2025
63. Sidney S. Nonparametric statistics for the behavioral sciences. J Nerv Ment Dis. 1957;125(3):497. https://journals.lww.com/jonmd/citation/1957/07000/NONPARAMETRIC_STATISTICS_FOR_THE_BEHAVIORAL.32.aspx. Accessed 1st July 2025.
64. Pearson KX. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Lond Edinb Dublin Philos Mag J Sci. 1900;50(302):157–75. https://doi.org/10.1080/14786440009463897.
65. Kruskal WH, and WAW. Use of Ranks in One-Criterion Variance Analysis. J Am Stat Assoc. 1952;47(260):583–621. https://doi.org/10.1080/01621459.1952.10483441.
66. Team TPD. Pandas: Python Data Analysis Library. https://pandas.pydata.org/. URL visited on 19th February 2025. Accessed 1st July 2025.
67. Community TS. SciPy: scientific computing library for Python. https://scipy.org/. URL visited on 19th February 2025. Accessed 1st July 2025.
68. Team TMD. Matplotlib: visualization with Python. https://matplotlib.org/. URL visited on 19th February 2025. Accessed 1st July 2025.
69. Developers TN. NumPy: The Fundamental Package for Scientific Computing with Python. https://numpy.org/. URL visited on 19th February 2025. Accessed 1st July 2025.
70. Sandve GK, Nekrutenko A, Taylor J, Hovig E. Ten simple rules for reproducible computational research. PLoS Comput Biol. 2013;9(10):e1003285. https://doi.org/10.1371/journal.pcbi.1003285.
71. Danchev V. Reproducible data science with Python: an open learning resource. J Open Source Educ. 2022. https://doi.org/10.21105/jose.00156.
72. PYPL. PopularitY of Programming Language for May 2025. 2025. https://pypl.github.io/PYPL.html. URL visited on 14th May. Accessed 1st July 2025.
73. TIOBE. TIOBE Index May 2025. 2025. https://www.tiobe.com/tiobe-index/. URL visited on 14th May. Accessed 1st July 2025.
74. Kaggle. State of data science and machine learning 2022. 2025. https://www.kaggle.com/kaggle-survey-2022. URL visited on 14th May. Accessed 1st July 2025.
75. Gibbons JD, Pratt JW. P-values: interpretation and methodology. Am Stat. 1975;29(1):20–5. https://doi.org/10.1080/00031305.1975.10479106.
76. Altman N, Krzywinski M. Interpreting P values. Nat Methods. 2017;14(3):213–4. https://doi.org/10.1038/nmeth.4210.
77. Goodman S. A dirty dozen: twelve P-value misconceptions. Semin Hematol. 2008;45(3):135–40. https://doi.org/10.1053/j.seminhematol.2008.04.003.
78. Andrade C. The P value and statistical significance: misunderstandings, explanations, challenges, and alternatives. Indian J Psychol Med. 2019;41(3):210–5. https://doi.org/10.4103/IJPSYM.IJPSYM_193_19.
79. Di Leo G, Sardanelli F. Statistical significance: p value, 0.05 threshold, and applications to radiomics—reasons for a conservative approach. Eur Radiol Exp. 2020;4(1). https://doi.org/10.1186/s41747-020-0145-y.
80. Wasserstein RL, Schirm AL, Lazar NA. Moving to a World Beyond "p < 0.05". Am Stat. 2019;73(sup1):1–19. https://doi.org/10.1080/00031305.2019.1583913.
81. Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD. The extent and consequences of *p*-hacking in science. PLoS Biol. 2015;13(3):e1002106. https://doi.org/10.1371/journal.pbio.1002106.
82. Gleason J. Comparative power of the ANOVA, randomization ANOVA, and Kruskal-Wallis Test [PhD Dissertation]. Wayne State University; 2013. https://digitalcommons.wayne.edu/oa_dissertations/658. Accessed 1st July 2025.
83. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). Biometrika. 1965;52(3–4):591–611. https://doi.org/10.1093/biomet/52.3-4.591.
84. Berger VW, Zhou Y. Kolmogorov-Smirnov test: overview. Wiley StatsRef: Statistics Reference Online. 2014. https://doi.org/10.1002/9781118445112.stat06558.
85. Ghasemi A, Zahediasl S. Normality tests for statistical analysis: a guide for non-statisticians. Int J Endocrinol Metab. 2012;10(2):486. https://doi.org/10.5812/ijem.3505.
86. Razali NM, Wah YB. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov Lilliefors and Anderson-Darling tests. J Stat Model Anal. 2011;2(1):21–33. http://www.de.ufpb.br/~ulisses/disciplinas/normality_tests_comparison.pdf. Accessed 1st July 2025.
87. Kikvidze Z, Moya-Laraño J. Unexpected failures of recommended tests in basic statistical analyses of ecological data. Web Ecol. 2008;8(1):67–73. https://doi.org/10.5194/we-8-67-2008.
88. Tang AS, Woldemariam SR, Miramontes S, Norgeot B, Oskotsky TT, Sirota M. Harnessing EHR data for health research. Nat Med. 2024;30(7):1847–55. https://doi.org/10.1038/s41591-024-03074-8.
89. Ma Y, Zheng J, Feng J, Chen L, Dong K, Xiao X. Neuroblastomas in eastern China: a retrospective series study of 275 cases in a regional center. PeerJ. 2018;6:e5665. https://doi.org/10.7717/peerj.5665.
90. Gucyetmez B, Atalan HK. C-reactive protein and hemogram parameters for the non-sepsis systemic inflammatory response syndrome and sepsis: what do they mean? PLoS ONE. 2016;11(2):e0148699. https://doi.org/10.1371/journal.pone.0148699.

91. Mollura M, Chicco D, Paglialonga A, Barbieri R. Identifying prognostic factors for survival in intensive care unit patients with SIRS or sepsis by machine learning analysis on electronic health records. PLoS Digit Health. 2024;3(3):e0000459. https://doi.org/10.1371/journal.pdig.0000459.
92. Jani BD, Mair FS, Roger VL, Weston SA, Jiang R, Chamberlain AM. Comorbid depression and heart failure: a community cohort study. PLoS ONE. 2016;11(6):e0158570. https://doi.org/10.1371/journal.pone.0158570.
93. Requena-Morales R, Palazón-Bru A, Rizo-Baeza MM, Adsuar-Quesada JM, Gil-Guillén VF, Cortés-Castell E. Mortality after out-of-hospital cardiac arrest in a Spanish region. PLoS ONE. 2017;12(4):e0175818. https://doi.org/10.1371/journal.pone.0175818.
94. Takashi Y, Ishizu M, Mori H, Miyashita K, Sakamoto F, Katakami N, et al. Circulating osteocalcin as a bone-derived hormone is inversely correlated with body fat in patients with type 1 diabetes. PLoS ONE. 2019;14(5):e0216416. https://doi.org/10.1371/journal.pone.0216416.
95. Cerono G, Chicco D. Ensemble machine learning reveals key features for diabetes duration from electronic health records. PeerJ Comput Sci. 2024;10:e1896. https://doi.org/10.7717/peerj-cs.1896.
96. Bartolucci D, Montemurro L, Raieli S, Lampis S, Pession A, Hrelia P, et al. MYCN impact on high-risk neuroblastoma: from diagnosis and prognosis to targeted treatment. Cancers. 2022;14(18):4421. https://doi.org/10.3390/cancers14184421.
97. Nakazawa A, Haga C, Ohira M, Okita H, Kamijo T, Nakagawara A. Correlation between the International Neuroblastoma Pathology Classification and genomic signature in neuroblastoma. Cancer Sci. 2015;106(6):766–71. https://doi.org/10.1111/cas.12665.
98. Bauer M, Gerlach H, Vogelmann T, Preissing F, Stiefel J, Adam D. Mortality in sepsis and septic shock in Europe, North America and Australia between 2009 and 2019-Results from a systematic review and meta-analysis. Crit Care. 2020;24(1):239. https://doi.org/10.1186/s13054-020-02950-2.
99. Beach SR, Januzzi JL, Mastromauro CA, Healy BC, Beale EE, Celano CM, et al. Patient Health Questionnaire-9 score and adverse cardiac outcomes in patients hospitalized for acute cardiac disease. J Psychosom Res. 2013;75(5):409–13. https://doi.org/10.1016/j.jpsychores.2013.08.001.
100. Ishii J, Nishikimi M, Kikutani K, Ohki S, Ota K, Anzai T, et al. Resuscitation attempt and outcomes in patients with asystole out-of-hospital cardiac arrest. JAMA Netw Open. 2024;7(11):e2445543. https://doi.org/10.1001/jamanetworkopen.2024.45543.
101. Bower JK, Meadows RJ, Foster MC, Foraker RE, Shoben AB. The association of percent body fat and lean mass with HbA1c in US adults. J Endocr Soc. 2017;1(6):600–8. https://doi.org/10.1210/js.2017-00046.
102. Hotelling H. The generalization of Student's ratio. Ann Math Stat. 1931;2(3):360–78. https://doi.org/10.1214/aoms/1177732979.

Publisher's Note