The American Journal of Pathology, Vol. 191, No. 5, May 2021

ELSEVIER

The American Journal of
PATHOLOGY

ajp.amjpathol.org

See related Commentary on page 805

REVIEW

Statistical Methods in Experimental Pathology

A Review and Primer

Douglas A. Mata* and Danny A. Milner, Jr[†‡]

From Foundation Medicine, Inc., Cambridge, Massachusetts; the American Society for Clinical Pathology,[†] Chicago, Illinois; and the Harvard T.H. Chan School of Public Health,[‡] Boston, Massachusetts*

Accepted for publication February 11, 2021.

Address correspondence to Douglas A. Mata, M.D., M.P.H., Foundation Medicine, Inc., 150 Second St., Cambridge, MA, 02141-2115; or Danny A. Milner, Jr, M.D., M.Sc., M.B.A., American Society for Clinical Pathology, 33 W. Monroe St., Ste. 1600, Chicago, IL 60605-5308. E-mail: dmata@foundationmedicine.com, douglas.a.mata@gmail.com, or dan.milner@ascp.org.

Correct use of statistical methods is important to ensure the reliability and value of the published experimental pathology literature. Considering increasing interest in the quality of statistical reporting in pathology, the statistical methods used in 10 recent issues of the *American Journal of Pathology* were reviewed. The statistical tests performed in the articles were summarized, with attention to their implications for contemporary pathology research and practice. Among the 195 articles identified, 93% reported using one or more statistical tests. Retrospective statistical review of the articles revealed several key findings. First, tests for normality were infrequently reported, and parametric hypothesis tests were overutilized. Second, studies reporting multisample hypothesis tests (eg, analysis of variance) infrequently performed *post hoc* tests to explore differences between study groups. Third, correlation, regression, and survival analysis techniques were underutilized. On the basis of these findings, a primer on relevant statistical concepts and tests is presented, including issues related to optimal study design, descriptive and comparative statistics, and regression, correlation, survival, and genetic data analysis. *(Am J Pathol 2021, 191: 784—794; https://doi.org/10.1016/j.ajpath.2021.02.009)*

Conducting experimental pathology research requires familiarity with a range of statistical concepts and proficiency in performing statistical tests. However, prior work has shown that pathologists report a poor overall understanding of statistics, suggesting that they would benefit from further training in epidemiologic study design and from mastering about a dozen commonly used statistical tests.[1]

In experimental pathology, the range of experiment types is vast and includes *in vitro* basic science studies, animal studies, qualitative and quantitative tissue morphology and morphometry analyses, and big data studies of genetic, transcriptomic, and proteomic data. Unfortunately, physicians and scientists with limited formal training in statistics often select incorrect analysis plans when preparing manuscripts for publication. Additionally even with an editor's focus on high-quality scientific data, many editorial boards and reviewers lack the expertise to provide sufficient statistical review to catch subtle errors in study design, test selection, and interpretation. Therefore, the scientific community, because of the sheer volume of articles published each year that experience these issues, may proceed down a path of scientific conclusion that is unjustified.

To ameliorate these issues, there has been a movement within the scientific community to increase the emphasis on statistical review at the grant and publication level.[2,3] This movement, although well intentioned, is hampered by a lack of individuals with statistical expertise and/or crossover knowledge to provide meaningful review of increasingly complex studies. A complementary approach is for scientific journals to periodically review the statistical tests used in their publications, observe what can be learned from these surveys, and provide rationale and support to future authors about how to make the right decisions for their studies.[4]

To that end, we surveyed a selection of articles published in *The American Journal of Pathology* over a 5y-period to

Disclosures: D.A.M. is an employee of Foundation Medicine, Inc., and owns stock in Roche Holdings.

https://doi.org/10.1016/j.ajpath.2021.02.009

understand the type and frequency of statistical tests performed. We then used these data as a starting point to discuss which tests a researcher should be familiar with as well as to make inferences about test utilization in the field of experimental pathology.

Survey of Articles

Ten issues of *The American Journal of Pathology* published between January 2014 and December 2018 were randomly selected for review. The articles published in the issues were examined to determine the type and number of statistical tests reported per article. Among the 10 issues, 195 articles were identified, 93% of which reported using one or more statistical tests. In all, 426 statistical tests were reported, with a median of two tests per article. A summary of the statistical tests reported by the articles is shown in Figure 1. These included tests of normality (eg, Kolmogorov-Smirnov test); methods of *P*-value correction for multiple testing (eg, Bonferroni); two-sample parametric (eg, *t*-test) and nonparametric (eg, Mann-Whitney-Wilcoxon) tests; multisample parametric (eg, parametric analysis of variance, including Tukey and Dunnett tests) and nonparametric (eg, Kruskal-Wallis) tests; *post hoc* tests (eg, Newman-Keuls); tests for assessing differences among categorical variables (eg, χ^2 and Fisher exact tests); correlation analysis (eg, Spearman and Pearson methods); generalized linear models (eg, linear regression); survival analysis (eg, Kaplan-Meier analysis and Cox proportional hazards regression); and other specialized methods, such as ΔCt analysis of real-time PCR data.

Retrospective review of the reported statistical tests revealed several key findings. Tests for normality, the formal method of determining whether a parametric or nonparametric test is indicated, were only reported in 5% of articles, even though many had small sample sizes (ie, <20 observations). Parametric tests, which assume that data are distributed according to a well-defined distribution and thus are generally applicable to studies with larger sample sizes, were accordingly overutilized, constituting 54% of tests. Two-sample parametric tests predominated, representing 32% of tests reported, followed by multisample parametric tests, representing 22% of tests reported. Only 8% of reported two-sample tests were nonparametric. Studies reporting multisample tests did not perform *post hoc* tests to explore differences between study groups 68% of the time. Few studies corrected *P* values for multiple testing. Lastly, correlation, regression, and survival analysis techniques were underused, even though several studies reported data that would have been amenable to analysis with such methods (Figure 1).

Taken together, these results indicate that there is room for improvement in the statistical methods of the published pathology literature. On the basis of these findings, a primer on relevant statistical concepts and tests is presented below, including issues pertaining to study design, descriptive statistics, normality assessment, two-sample and multisample comparisons, regression and correlation analysis, categorical data analysis, survival analysis, and genetic data analysis.

Selecting an Appropriate Study Design

High-quality research in pathology should ideally begin with rigorous attention to study design and statistical methods.[5] However, the skills needed to pursue these objectives are not often taught in pathology training programs, unlike in training programs in data science. For example, in training epidemiologists and biostatisticians, emphasis is placed on experimental design before beginning data collection. This process involves generating an a priori statistical analysis plan, including anticipated data types, sample size calculations, and test selection. In only rare instances does the researcher designing an experiment not know the form the data will take; therefore, the generation of a statistical analysis plan, which should guide

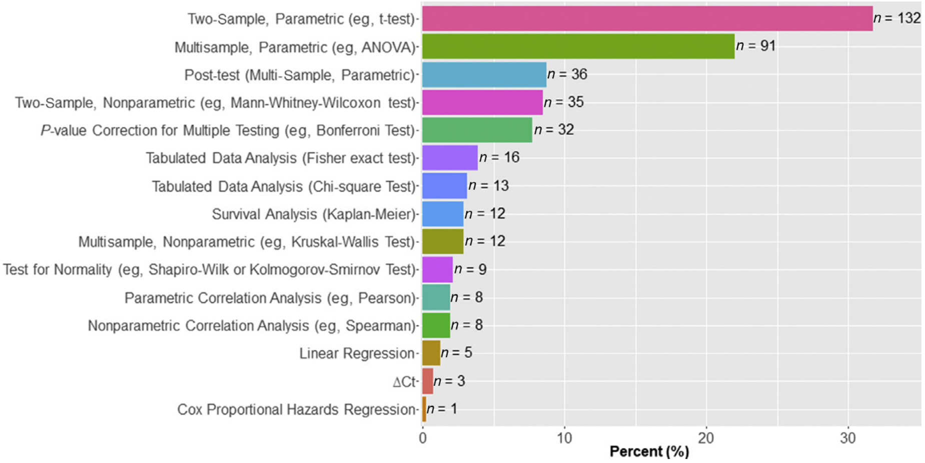

Figure 1 Statistical tests utilized by the articles surveyed. A bar chart is shown with the breakdown of test type, total number encountered, and percentage of total. The graph shows all tests (not all articles). $n = 426$. ANOVA, analysis of variance.

experiments just as the technical protocol does, should be considered mandatory.

Selecting a proper study design should coincide with the process of generating a research question. Turning an idea for an experimental pathology study into an answerable question is an important part of the process. A good question is FINER—feasible, interesting, novel, ethical, and relevant—and should be written up as a formal scientific proposal with a literature review, hypothesis, and proposed methods before commencing data collection.[6] Completing this exercise will inform study design selection—be it case-control, cross-sectional, cohort, randomized, or something else—and the approach to the data analysis. In experimental pathology, cross-sectional or retrospective cohort studies are common study designs.

For any proposed experimental pathology study, one should read the relevant guidelines before collecting data (Table 1). The EQUATOR Network (Enhancing the QUAlity and Transparency Of health Research) provides resources for reporting research studies, providing guidelines for all standard epidemiologic study designs as well as for basic science and animal studies (*http://www.equator-network.org*, last accessed February 17, 2021).[7] Although these guidelines primarily advise how to report results, they also provide an implicit framework for how studies should be conducted. Of particular relevance to pathology are the Standard Protocol Items: Recommendations for Interventional Trials statement, which discusses how to prepare an a priori study protocol before data collection[8]; the Animal Research: Reporting of *In Vivo* Experiments guidelines, which discuss how to report *in vivo* animal experiments[9]; the Case Reports guidelines, which provide instructions on how to prepare case reports[10]; and the Strengthening the Reporting of Observational Studies in Epidemiology guidelines, which cover the reporting of observational studies such as cross-sectional and cohort studies.[11] The Consolidated Standards of Reporting Trials statement, which advises on reporting parallel group randomized trials, is also of special relevance to experimental pathology because its key concepts, although intended for humans, can be applied to experimental intervention studies involving animals.[12]

Statistical Software

Once a research question, study design, and statistical analysis protocol have been established, data collection and analysis can commence. Familiarity with one or more commonly used data management systems and statistical software packages is essential. Although beginners might make use of free web-based analysis tools, such as MedCalc (MedCalc Software Ltd, Ostend, Belgium) and OpenEpi (*http://www.openepi.com*, last accessed February 17, 2021), intermediate users will likely be most comfortable with Excel (Microsoft Corp., Redmond, WA), which not only allows for data entry, storage, and filtering, but also provides a graphical user interface for graphical data exploration, descriptive and comparative statistics, and correlation and regression analysis through the Analysis ToolPak. Other easily accessible graphical user interface–based options include GraphPad Prism (GraphPad Software, San Diego, CA), Minitab (Minitab, LLC, State College, PA), SPSS (IBM Corp., Armonk, NY), and PSPP (*http://www.gnu.org/software/pspp*, last accessed February 17, 2021). However, most researchers will want to develop expertise in a robust statistical programming environment, such as R (The R Foundation for Statistical Computing, Vienna, Austria), Stata (StataCorp LLC, College Station, TX), SAS (SAS Institute Inc., Cary, NC), or Anaconda/Python (Anaconda Inc., Austin, TX). The following sections assume the reader has access to one or more of these tools.

Data Exploration and Descriptive Statistics

In pathology, many types of studies include observations that need to be described—and have value—but do not actually involve comparisons requiring a statistical test.

Table 1 Reporting Guidelines Relevant to Experimental Pathology Research

Type	Example	Description
Study protocols	SPIRIT statement	Guidelines for preparing *a priori* study protocols; emphasis on clinical trials, but relevant for other study types.
Preclinical animal studies	ARRIVE guidelines	Reporting guidelines for any subfield of biomedical research that utilizes laboratory animals.
Case reports	CARE guidelines	Consensus guidelines on data collection and chart review for clinical case reports.
Observational studies	STROBE statement	Guidelines for reporting case-control, cross-sectional, and cohort studies; broadly relevant to all pathology research.
Randomized trials	CONSORT statement	Guidelines for reporting parallel-group randomized trials.
Diagnostic accuracy studies	STARD guidelines	Standards for reporting the accuracy of medical diagnostic tests.

The Equator Network provides >400 other reporting guidelines with broad relevance across health research (*http://equator-network.org*). Authors are encouraged to search for reporting guidelines relevant to their study types before commencing data collection and analysis.

ARRIVE, Animal Research: Reporting of *In Vivo* Experiments; CARE, Case Reports; CONSORT, Consolidated Standards of Reporting Trials; SPIRIT, Standard Protocol Items: Recommendations for Interventional Trials; STARD, Standards for Reporting of Diagnostic Accuracy; STROBE, Strengthening the Reporting of Observational Studies in Epidemiology.

Researchers may encounter new entities (eg, tumors or disease states) or note an important phenomenon within a commonly used experimental system that should be accounted for and described (eg, a preponderance of *Mycoplasma* in cell cultures). The methods of descriptive statistics provide a means to summarize a data set using measures of frequency, central tendency, dispersion or variation, and position or rank.[13] These methods contrast with those of inferential statistics, which involve the use of hypothesis tests to generalize the findings from a smaller data set to a larger population. More importantly, the nature of the observations themselves often dictates the appropriate statistical description needed, and an analysis plan should ideally be prespecified before beginning a descriptive study, just as one would be for a randomized trial.

Regardless of the study design, the first step after data collection should be description rather than comparison. Once data have been collected, each variable in the data set should be visually inspected, explored, and summarized in accordance with the prespecified analytical plan defined before data collection commenced. Steps may include, in addition to providing textual descriptions of phenomena, i) inspecting data for typos and missing values; ii) calculating measures of central tendency (eg, mean, median, and mode) and dispersion (eg, SD, range, and interquartile range) of each continuous variable; iii) plotting continuous variables as scatter plots, histograms, box-and-whisker plots, or violin plots to visually assess for normality and potential outliers; and iv) tabulating categorical variables and determining the percentage of observations within each category.

Visualizing continuous variables using a histogram allows the researcher to determine by visual inspection whether the variables are normally distributed, which may inform the type of hypothesis testing to be performed (Figure 2). Notably, data that are not normally distributed, such as the right-skewed example in Figure 2, can often be transformed to a normal gaussian distribution by way of logarithmic or square root transformation.[14] Assessing for normality is a key component of the data exploration process because differences among normally distributed variables can be assessed with parametric hypothesis tests, whereas those among nonnormally distributed variables are better assessed with nonparametric tests.[15] However, it is important to remember that nonparametric tests can still be used on normally distributed data; in fact, nonparametric tests are likely to detect meaningful between-group differences and less likely to detect statistically significant differences that are in actuality clinically insignificant.

In addition to visual inspection, statistical tests of normality, such as the D'Agostino K-squared test, the Kolmogorov-Smirnov test, and the Shapiro-Wilk test are available, which test the data against the null hypothesis that they are normally distributed. Each of these tests has different performance characteristics, and none is particularly powerful for small sample sizes. They may not reject the null hypothesis of normality, even for data that deviate significantly from the gaussian distribution. For small sample sizes, we therefore recommend that nonparametric tests be used even if the data appear normally distributed on formal testing. One should also note that when sample sizes are exceptionally large, formal testing may reject the null hypothesis of normality, even for trivial deviations that have little to no effect on the conclusions drawn from parametric hypothesis tests. Therefore, for most medical data sets, visual inspection for normality is sufficient.

Regardless of whether visual inspection or formal statistical testing of normality is conducted, prespecifying the rationale for the decision and the downstream statistical analysis plan before conducting experiments and making further inferences is key. There is a significant risk of bias associated with defining a statistical testing approach after detailed examination of the data.

Parametric and Nonparametric Two-Sample and Multisample Comparisons

Once a comparison is introduced to an experiment, statistical tests are required to determine whether observed differences are likely to represent real differences. The methods of inferential statistics involve the use of statistical hypothesis tests to generalize findings from a sample to the larger population from which the sample was drawn. Choosing the appropriate comparative test depends on the type of data to be compared. A key distinction is between numeric and categorical variables. Numeric variables are either described with discrete values (eg, count data, such as the number of positive blood cultures in a patient hospitalized with sepsis or the number of mitoses per high-power field in a gastrointestinal stromal tumor) or take on a continuum of values (eg, age at initial histopathologic diagnosis of glioblastoma or the variant allele fraction of a mutant *NF1* allele by next-generation sequencing).

If a numeric-dependent variable is normally distributed, then a parametric statistical test can be used to assess whether it differs between two or more groups.[16] For example, a two-sample *t*-test could be used to determine whether age at diagnosis for prostatic adenocarcinoma differed according to whether patients were Black or White. For simultaneously comparing age at diagnosis among patients who were Asian, Hispanic, or Pacific Islander, a multisample comparison could be performed using analysis of variance, and if a significant difference was noted, a *post hoc* test could suggest which racial group was more likely to be diagnosed at a younger or older age. For multisample tests, either parametric or nonparametric, *post hoc* tests can be used to investigate the significance of an individual effect size.[17] However, a caveat is that studies are rarely adequately powered for definitive *post hoc* testing. The results of *post hoc* testing should generally be considered exploratory and ideally subjected to independent assessment in a separate experiment. Furthermore, researchers should

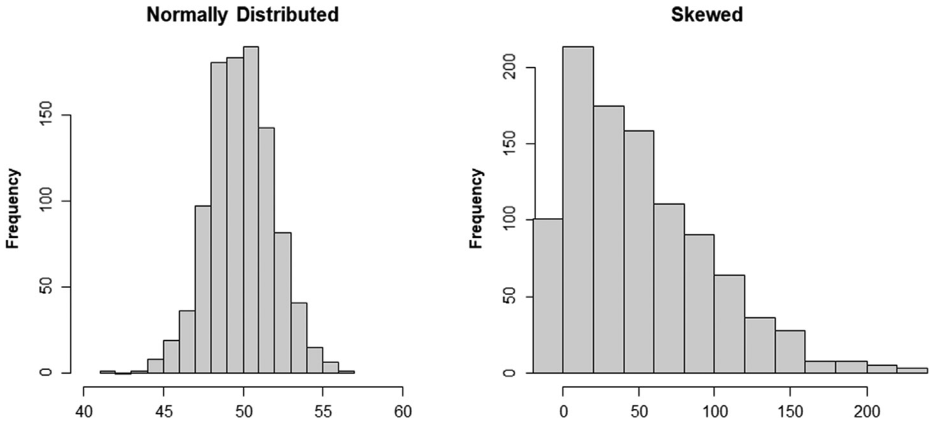

Figure 2 Normally distributed data compared with skewed data. **Left panel:** Visual inspection of the data reveals them to be normally distributed. The median and mean are both 50, and the Shapiro-Wilk normality test yields a *P* value of 0.21, confirming this impression. **Right panel:** Visual inspection of the data reveals a right skew. The median and mean are discordant (77 versus 50), and the Shapiro-Wilk test yields a statistically significant *P* value of <0.001, confirming the data are not normally distributed.

not use *post hoc* tests to data dredge subgroups if their original hypothesis was not supported by their experiment.

More importantly, because sample sizes in pathology are often small, nonparametric tests should be used at least as often as parametric tests, if not more often.[18] Nonparametric tests are used for data with a skewed distribution (eg, tumor mutational burden in glioblastoma patients with and without temozolomide-induced hypermutation) or taking the form of a discrete or ranked scale (eg, a Likert scale).[19] These tests have the advantage of relaxing assumptions about the distribution of the data. They work by assigning ranks to data and comparing the ranks without accounting for the actual size difference between sample values. Parametric tests, like *t*-tests, are easily influenced by extreme outlier values. In contrast, nonparametric tests are not hindered by outliers, as the rank scheme between the values remains unchanged. It bears repeating, a nonparametric test could also be applied to normally distributed data (eg, the Mann-Whitney-Wilcoxon test could substitute for the *t*-test in the prostate cancer example above); although this may slightly reduce power to detect differences, truly meaningful differences would still likely be detected.

Linear and Logistic Regression Analysis

Regression analysis is a principal tool for determining the behavior of a dependent outcome variable [eg, such as oxygen saturation measured by pulse oximetry in a patient with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection] in relation to one or more independent predictor variables (eg, patient age, smoking history, or number of comorbid medical conditions). A classic example in research is measuring one or more continuous or categorical independent (predictor) variables and using linear regression to determine their effect on a continuous dependent (outcome) variable of interest.

Example: Age, sex, race, history of diabetes, weight, body mass index, number of years smoking cigarettes, and average number of cigarettes smoked per day and patient serum low-density lipoprotein (LDL) levels, provide a set of nine variables. Of these, age, weight, body mass index, and number of years smoking are continuous variables (ie, they can have any value, and math can be performed on them). Sex, race, and history of diabetes are categorical variables. The average number of cigarettes smoked per day is an integral variable greater than zero so it may be evaluated as an ordinal or a categorical (collapsed) variable. Our hypothesis is that LDL concentration is related to and predicted by one or more of the collected variables. Note that, given the nine variables, any of them can be used as the dependent (outcome) variable to examine how others relate to it. However, in practice, a scientific hypothesis is being tested and so a direction and purpose to the questions should be obvious. Because LDL cholesterol concentration is a continuous variable, linear regression is a good choice for determining whether age, sex, race, or any of the other collected predictor variables are associated with an increase or decrease in LDL levels. The magnitude and direction of each association would be summarized by the slope of the fitted line between the outcome (ie, LDL) and predictor variables.

Logistic regression is similar to linear regression but has a dependent variable that is most commonly binary (eg, present versus absent).[20] For example, whether a patient infected with SARS-CoV-2 develops pneumonia on chest X-ray could be represented as a binary dependent (outcome) variable (yes versus no), and one could fit a logistic regression model to the data to determine whether the independent (predictor) variables age, smoking history, or number of comorbid medical conditions were associated with an increased risk of pneumonia. The strength of the association between each predictor and the outcome is represented by the odds ratio (ie, the ratio of the probability that the outcome occurs/the probability that it does not occur).[21]

Like all statistical models, linear and logistic regression make certain assumptions about the distribution of data that should be assessed before model building. Linear regression assumes that the relationship between predictor and outcome variables is linear. It also assumes homoscedasticity (ie, that the noise in the relationship between the predictor and outcome variables is the same across all values of the predictor variables) and that input variables are normally distributed. Logistic regression also assumes linearity [ie, linearity between logit(p) or log (p/1-p) and outcome]. Both methods also assume low or no collinearity among independent variables. For example, one should not include both hemoglobin and hematocrit concentration in the same regression model, in part because both variables provide the model with essentially identical information. One can ensure that potential predictor variables are not unduly collinear with one another by performing correlation analysis before model building.

An important step in regression modeling is determining the number of predictor variables that are permissible to include in a multivariate regression model. A common rule of thumb is that one predictor can be added for every 10 events in the data set.[22] In the LDL cholesterol example, there were eight predictor variables and one outcome variable. Therefore, at least 80 patients would be needed in the data set if all eight predictors were to be included in a linear regression model to predict LDL cholesterol. However, if the same data set were to be used to predict the risk of acute myocardial infarction (coded as present versus absent) using logistic regression and only 20 patients experienced a heart attack, then it would have been possible to include only two predictors in the model. Notably, the 1 in 10 rule is conservative, and, in some instances, it is possible to justify including additional predictors.[23]

A separate issue is deciding which predictor variables to include in a regression model in the first place. In general, one should first build univariate regression models for each potential predictor variable to obtain unadjusted estimates.[11] Then, when building multivariate models, predictor variables should ideally be selected in a rational, hypothesis-driven manner based on the understanding of disease pathophysiology. For example, body mass index and smoking status are typically adjusted for in multivariate models of coronary risk because prior research has established them as biologically important risk factors that may confound the observed univariate associations of other potential predictor variables with coronary outcomes. Using stepwise regression—a computer-aided method that selects predictor variables by automatic procedure is typically not recommended to select predictor variables for model inclusion. Stepwise regression is widely regarded as a form of data dredging, although it may occasionally be useful for exploratory hypothesis generation.[24]

Lastly, although linear and logistic regression analysis are commonly used in medical research, other variations of regression analysis—such as polynomial, quantile, Poisson, and Cox regression, the latter of which is discussed below—may be better suited for some study questions. Researchers should seek formal statistical consultation when planning their studies to ensure that the most appropriate methods are selected before data collection and analysis.

Correlation Analysis

Correlation analysis is similar to regression in that it examines the statistical relationship between two variables.[25] However, unlike regression, it is not generally used for prediction and does not assume a causal relationship.[26] Rather, it seeks to identify and quantify the degree of association or relatedness between two variables (eg, hemoglobin and hematocrit), which is summarized as a correlation coefficient. As for other statistical techniques, multiple types of correlation analysis exist, and the appropriate method depends on the structure of the data of interest. Methods of correlation analysis fall into two categories: linear correlation and nonparametric correlation. Continuous variables (eg, LDL and total cholesterol levels) can be analyzed using the Pearson product-moment correlation coefficient, a linear correlation method that produces a value between -1 and 1, summarizing both the strength and the direction of the correlation. This method generally provides a more valid result if the data are normally distributed; also, it is not particularly robust to outliers. Relationships between ordinal variables (eg, the experience level of an anatomic pathologist and the Gleason grade assigned to a prostate core biopsy) can be assessed using the nonparametric Spearman or Kendall rank correlation coefficients. These methods can also be used for continuous variables regardless of whether they are normally distributed. As their names imply, the rank correlation coefficients quantify the degree of similarity between two rankings and thus are more robust to outliers. Before performing regression analysis, generating a correlation matrix of all variables in a data set can be an informative way to discover relationships between variables as well as identify collinearity.

Analysis of Categorical Data

Binary categorical (eg, the presence of absence of a disease state) and multilevel categorical (eg, the pathologic stage of a tumor) data are commonly encountered in pathology. These data are often summarized in tabular form for analysis. As in case for numeric data, it is often of interest to conduct two-sample or multisample comparisons involving variables that are categorical. For example, the histologic diagnosis of a brain tumor (eg, glioblastoma or medulloblastoma) can be tabulated against the location of the tumor (eg, frontal, parietal, temporal, occipital, or cerebellar) in a 2×5 table, and if the sample were large (eg, greater than five observations per cell of the table), a χ^2 hypothesis test can be performed to assess for a statistically significant difference in neuroanatomic distribution between the tumor types. For small samples, the Fisher exact test can be used instead. Notably, the χ^2 and Fisher exact tests are considered nonparametric tests; both can be used for multisample comparisons. As for numeric data, it is also often of interest to examine the correlation between categorical variables. Tabulated data are therefore also important in assessing interrater agreement (eg, determining whether two pathologists interpret an immunohistochemical stain in the same manner, often summarized using the Cohen κ).

Data tabulation is also used in the evaluation of diagnostic test accuracy (eg, for calculating sensitivity, specificity, and positive and negative predictive values). Sensitivity and specificity, the probabilities that a test result will be positive or negative when the disease is present or absent, respectfully, are intrinsic to a diagnostic test and do not depend on disease prevalence. The relationship between the sensitivity and specificity of a diagnostic test can be explored through a receiver operating characteristic curve, which also provides a visual assessment of diagnostic test accuracy (for a helpful online calculator, see *http://rad.jhmi.edu/jeng/javarad/roc/JROCFITi.html*, last accessed February 17, 2020). Conversely, positive and negative predictive values, the probabilities that the disease is present or absent when the test is positive or negative, respectfully, depend on the prevalence of the disease in the population being tested. These prevalence-dependent values are of immense practical importance and are often overlooked in pathology studies (eg, in studies of immunohistochemical markers for diagnosis of specific malignancies).

Example: A commercial nucleic acid amplification test is implemented for SARS-CoV-2 to use as a pre-admission screening tool for patients at a hospital. The package insert reports seemingly excellent performance characteristics, including a sensitivity of 91.0% and a specificity of 99.0%. A clinician orders the test on an otherwise healthy patient being evaluated in the emergency room for a fractured ankle, and the test result is positive. The reported prevalence of SARS-CoV-2 infection in the community is approximately 1%. Using this information, the clinician is told that the positive predictive value—that is, the probability that the disease is present when the test is positive—is only 50.0% for this patient. The patient is tested again the following day, and the result is negative. The clinician then tests a second patient, an elderly man with fever, shortness of breath, anosmia, and diarrhea. The prevalence of SARS-CoV-2 infection in patients with this tetrad of symptoms in the community is >50%. The positive predictive value for this patient is 100%. Repeated diagnostic testing is not necessary, and the patient is admitted and treated. These examples illustrate the importance of considering disease prevalence and clinical suspicion (ie, the pretest probability) in the statistical evaluation of diagnostic tests.

More importantly, researchers should approach studies involving diagnostic tests with the same rigor that they would a randomized trial. The Standards for Reporting of Diagnostic Accuracy Guidelines provide a list of essential items to consider when designing, analyzing, and reporting studies of diagnostic accuracy.[27] Readers should also be aware of useful online statistical calculators that helpfully explain the statistical concepts behind these tests (*http://medcalc.org/calc/diagnostic_test.php* and *http://kennis-research.shinyapps.io/Bayes-App*, both accessed February 17, 2021).

Survival Analysis

Survival analysis refers to the suite of statistical techniques used to analyze time-to-event data.[28] Examples in pathology include overall survival (eg, time from diagnostic biopsy of a neoplasm until death attributable to any cause), cancer-specific survival (eg, time from diagnostic biopsy until death attributable to the cancer with which the patient or experimental animal was diagnosed), and progression-free survival (eg, the time from treatment assignment in an intervention trial to disease progression, the definition of which depends on the illness being studied). Time-to-event data are a composite of a continuous variable (ie, time, usually measured in months or years) and a binary variable (ie, the event, such as whether one is deceased or alive). For subjects who have died, the time interval is calculated from diagnosis until the date of death, and the event (death) is represented in the data set by the numeral 1. For subjects who are alive, the time interval is calculated from the date of diagnosis until the last known date at which they were alive (eg, the last time they were seen in clinic) or known to be progression-free (eg, the date of their last magnetic resonance imaging). Their status is represented by the numeral 0, indicating that their outcome is censored.

Together, these variables can be input into a survival function, which examines probability that the subject will survive (or not progress) beyond a specified time. In the simplest case, the two variables (time and status) can be used to plot a Kaplan-Meier survival curve, a nonparametric function that plots the percentage survival on the y axis and time on the x axis. Each time an event occurs, the curve steps down, generating a line resembling a ladder. Censored subjects are represented by tick marks. An example of this approach could be a study of transgenic mice with experimentally induced glioblastoma. One could generate a Kaplan-Meier curve of all mice to estimate median survival time in the entire cohort, finding that mice live on average 150 days after diagnosis. To compare groups stratified by another variable (eg, *NF1* knockdown), one could generate a third categorical variable that lists *NF1* knockdown status as present or absent for each mouse. Then, plotting both groups on the same curve, one could ascertain that mice with *NF1* knockdown live only 100 days compared with 200 days on average for those with intact *NF1* signaling (Figure 3). A formal statistical test, the log-rank test, can be used to generate a P value and determine whether the difference is statistically significant. The clinical significance also involves one's judgment as a scientist or physician.

An extension of the Kaplan-Meier concept is the proportional hazards model, the most common of which is known as Cox proportional hazards regression.[29,30] The Cox model produces a hazard ratio, that is, the ratio of the hazard of the event (eg, death) according to whatever predictor variable one is interested in (eg, age, sex, or *NF1* knockdown status). For example, one would expect a subject's hazard of death to increase with each 1-month increase in age (a continuous variable). The hazard of death might also increase with each 1-mm increase in radiologic tumor size at diagnosis. Or, being male rather than female (a binary categorical variable) might be associated with an increased hazard of death. In Cox regression, one uses the same outcome as in a Kaplan-Meier model (namely, a composite of time and status). In the case of univariate Cox regression of a binary predictor, one would reach the same conclusion as a Kaplan-Meier model. For example, one could use *NF1* knockdown status (present versus absent) as a predictor and overall survival as an outcome, and the resulting magnitude of the hazard ratio and *P* value would be concordant with that from Kaplan-Meier analysis. A useful aspect of Cox regression, however, is that one can more easily perform a multivariable analysis and add potential confounders (eg, age, sex, mouse strain, or other pathologic or genetic variables) to the model, accounting for their effects on the association of interest.

Like linear and logistic regression, Cox proportional hazards regression makes certain assumptions about the distribution of input data that should be considered before building a model. Chief among these is the assumption of proportional hazards, that the ratio of the hazards for any two study subjects is constant as time passes. For example, in a study of relapse-free survival in prostate cancer patients receiving a placebo pill or an investigational drug, if a patient receiving the placebo has a risk of relapse at an initial time point that is twice as high as that of another patient receiving the drug, and at all later times the risk of relapse remains twice as high, then the assumption of proportional hazards is met. Whether this assumption is met by a Cox model can be tested by examining the Schoenfeld residuals. However, in real life, the assumption of proportional hazards is rarely met in medical studies. When hazards are nonproportional, the predictor variable in the model interacts with time. For example, screening smokers with low-dose chest computed tomography scans may have an immediate effect on identifying occult lung nodules (leading to an elevated hazard ratio of developing lung cancer in the month after the scan) and have a delayed preventive effect (and thus lower hazard) of being diagnosed with lung cancer years down the road (because low-stage lung cancers detected by screening were surgically resected). More importantly, inferring more than the direction of an effect from a hazard ratio is conceptually difficult; therefore, researchers should always provide measures of survival differences to give additional context to studies involving time-to-event data.

Analysis of Genetic Data

With the advent of next-generation sequencing and high-resolution microarrays, an increasing number of experimental pathology studies are incorporating not only traditional clinical, histologic, and immunohistochemical measures, but also high-dimensional genetic

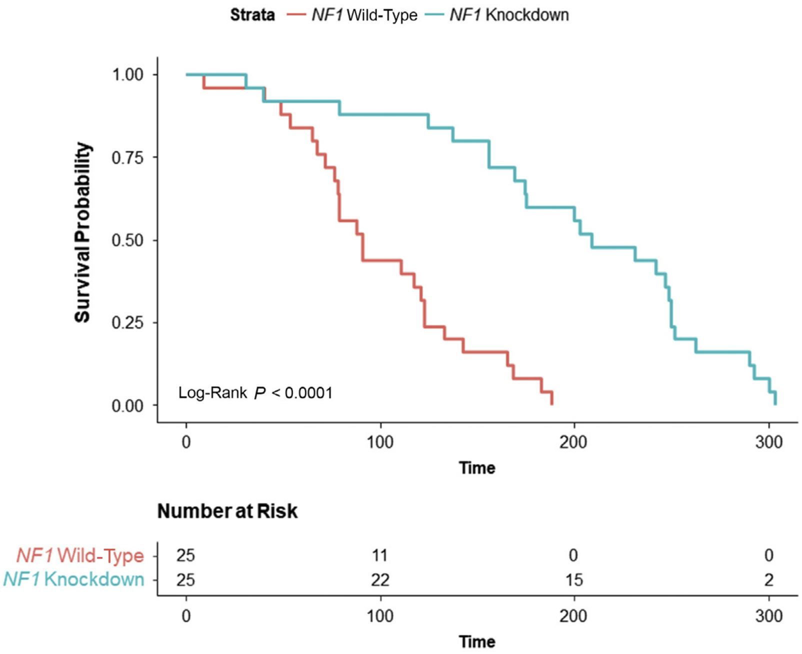

Figure 3 Example Kaplan-Meier survival curve of mice with experimentally induced glioblastoma with and without *NF1* knockdown. The dummy data used in this example were generated using a random sampling function in the statistical computing program R. The nonparametric log-rank test was used to assess for a difference between the two survival distributions.

and epigenetic data. For example, the Memorial Sloan Kettering Cancer Center Integrated Mutation Profiling of Actionable Cancer Targets next-generation DNA sequencing platform targets 468 genes and select introns to produce data on single-nucleotide variants, small insertions and deletions, copy number variations, and structural variants.[31] Similar data can be gleaned from the Brigham and Women's Hospital OncoPanel assay,[32] the Foundation Medicine FoundationOneCDx assay,[33] and the Tempus Labs xT assay.[34] For diseases such as brain tumors and sarcomas, methylation analysis can provide important diagnostic information.[35] For example, the Infinium MethylationEPIC assay provides *MGMT* promoter hypermethylation status as well as data on >850,000 CpG methylation sites across the genome.[36,37]

Although the magnitude of data produced by these technologies may sound daunting, it is important to remember that all the familiar tests described above can be applied to these data. For example, one could sequence 50 lung tumors to assess for the presence of an oncogenic *EGFR* alteration, which could be coded in binary (eg, present or absent). This variable could be combined with histologic diagnosis (eg, adenocarcinoma versus squamous cell carcinoma) in a 2 × 2 table, from which one could calculate an odds ratio, a CI, and a Fisher exact test *P* value to determine the extent to which oncogenic *EGFR* alterations are more common in lung adenocarcinomas rather than squamous cell carcinomas. Or, the $\log_2$ ratio (ie, the fold change) between the prevalence of the genetic alteration in the adenocarcinoma compared with the squamous cell carcinoma group could be calculated. In a Foundation Medicine data set with 664 interrogated genes, this could be performed on a per-gene basis and plotted in a volcano plot, a special type of scatter plot that allows the viewer to visually determine which genetic alterations are enriched for in a particular group of patients (Figure 4). When performing multiple hypothesis tests in a data set that experiences the small n, large p problem (ie, few cases but many potential predictor variables), correction for multiple testing, as by a Bonferroni correction or the Hochberg sequential procedure, is essential.[38] Multiple online graphical user interface−based tools are available to simplify the analysis of genetic data, such as cBioPortal[39] and PathwayMapper.[40]

Summary

The results of this survey of articles from *The American Journal of Pathology* indicate that there is room for improvement in statistical methods. Because experimental sample sizes in pathology studies are often small, tests for normality are essential, and nonparametric tests should be used to assess for potential differences between variables. Familiarity with a wider array of statistical techniques, including multisample comparisons, regression and correlation analysis, survival analysis, and high-dimensional genetic data analysis, can help experimental pathologists make

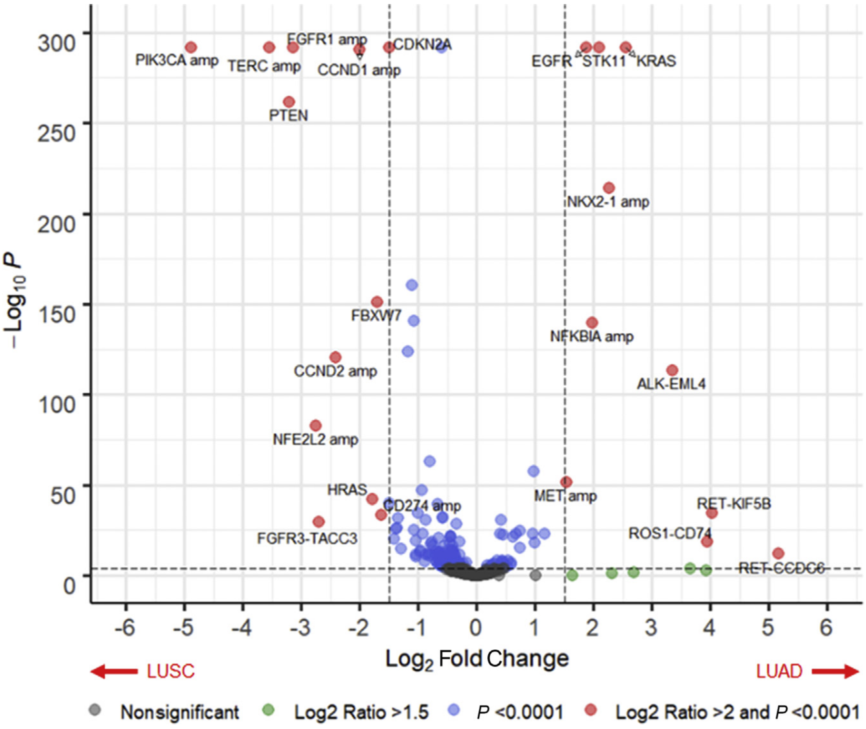

Figure 4 Volcano plot of gene mutation frequencies in lung adenocarcinoma (LUAD) versus lung squamous cell carcinoma (LUSC). The *y* axis shows the $-\log_{10}$ of the *P* value, whereas the *x* axis shows the $\log_2$ of the ratio between the prevalence of the specified gene alteration in 55,644 cases of LUAD versus 13,297 cases of LUSC. The *P*-value cutoff is represented by the **horizontal dashed line**. The **vertical dashed lines** denote $\log_2$ fold changes of −1.5 and 1.5, respectively. The analysis confirms that cases of LUAD are relatively enriched for *KRAS*, *EGFR*, and *STK11* mutations; *NKX2.1*, *NFKBIA*, and *MET* amplifications (amp); and *ALK*, *RET*, and *ROS1* fusions. Conversely, cases of LUSC are relatively enriched for *CDKN2A* and *PTEN* inactivating mutations, *FBXW7* mutations, FGF cluster and *CD274* (PD-L1) amplifications, and *FGFR3-TACC3* fusions, among other alterations. Because 664 hypothesis tests were performed, a Bonferroni correction was performed to determine the *P*-value cutoff: 0.05/664 = 0.0001. These real-world patient data are from the FoundationCore comprehensive genomic profiling database (*http://insights.foundationmedicine.com*, last accessed February 17, 2021, registration required).

the most of their studies to improve research findings and patient care.

References

1. Schmidt RL, Chute DJ, Colbert-Getz JM, Firpo-Betancourt A, James DS, Karp JK, Miller DC, Milner DA, Smock KJ, Sutton AT, Walker BS, White KL, Wilson AR, Wojcik EM, Yared MA, Factor RE: Statistical literacy among academic pathologists: a survey study to gauge knowledge of frequently used statistical tests among trainees and faculty. Arch Pathol Lab Med 2017, 141:279−287
2. Altman DG: Statistical reviewing for medical journals. Stat Med 1998, 17:2661−2674
3. Goodman SN, Altman DG, George SL: Statistical reviewing policies of medical journals: caveat lector? J Gen Intern Med 1998, 13: 753−756
4. Arnold LD, Braganza M, Salih R, Colditz GA: Statistical trends in the Journal of the American Medical Association and implications for training across the continuum of medical education. PLoS One 2013, 8:e77301
5. Milner DA, Meserve EEK, Soong TR, Mata DA: Statistics for Pathologists. ed 1. New York, NY, Demos Medical Publishing, 2016
6. Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB: Designing Clinical Research. ed 4. Philadelphia, PA, LWW, 2013
7. Altman DG, Simera I: A history of the evolution of guidelines for reporting medical research: the long road to the EQUATOR Network. J R Soc Med 2016, 109:67−77
8. Chan A-W, Tetzlaff JM, Altman DG, Laupacis A, Gøtzsche PC, Krleža-Jerić K, Hróbjartsson A, Mann H, Dickersin K, Berlin JA, Doré CJ, Parulekar WR, Summerskill WSM, Groves T, Schulz KF, Sox HC, Rockhold FW, Rennie D, Moher D: SPIRIT 2013 statement: defining standard protocol items for clinical trials. Ann Intern Med 2013, 158:200−207
9. Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG: Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. PLoS Biol 2010, 8:e1000412
10. Gagnier JJ, Kienle G, Altman DG, Moher D, Sox H, Riley D; CARE Group: The CARE guidelines: consensus-based clinical case report guideline development. J Clin Epidemiol 2014, 67:46−51
11. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative: The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 2007, 370:1453−1457
12. Schulz KF, Altman DG, Moher D; CONSORT Group: CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. Trials 2010, 11:32
13. Porta M: A Dictionary of Epidemiology. ed 6. Oxford, Oxford University Press, 2014
14. Feng C, Wang H, Lu N, Tu XM: Log transformation: application and interpretation in biomedical research. Stat Med 2013, 32:230−239
15. Mishra P, Pandey CM, Singh U, Gupta A, Sahu C, Keshri A: Descriptive statistics and normality tests for statistical data. Ann Card Anaesth 2019, 22:67−72
16. Geisser S, Johnson WM: Modes of Parametric Statistical Inference. ed 1. Hoboken, NJ, Wiley-Interscience, 2006
17. Tukey J: Comparing individual means in the analysis of variance. Biometrics 1949, 5:99−114
18. Altman DG, Bland JM: Parametric v non-parametric methods for data analysis. BMJ 2009, 338:a3167
19. Lachin JM: Nonparametric statistical analysis. JAMA 2020, 323: 2080−2081
20. Tolles J, Meurer WJ: Logistic regression: relating patient characteristics to outcomes. JAMA 2016, 316:533−534
21. Norton EC, Dowd BE, Maciejewski ML: Odds ratios—current best practice and use. JAMA 2018, 320:84−85
22. Harrell FE, Lee KL, Mark DB: Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996, 15:361−387
23. Vittinghoff E, McCulloch CE: Relaxing the rule of ten events per variable in logistic and Cox regression. Am J Epidemiol 2007, 165: 710−718
24. Livingston E, Cao J, Dimick JB: Tread carefully with stepwise regression. Arch Surg 2010, 145:1039−1040
25. Rodgers JL, Nicewander WA: Thirteen ways to look at the correlation coefficient: In: The American Statistician, 42. Taylor & Francis, 1988. pp. 59−66
26. Asuero AG, Sayago A, González AG: The correlation coefficient: an overview: In: Critical Reviews in Analytical Chemistry, 36. Milon Park, Oxfordshire, UK: Taylor & Francis, 2006. pp. 41−59
27. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, Lijmer JG, Moher D, Rennie D, de Vet HCW, Kressel HY, Rifai N, Golub RM, Altman DG, Hooft L, Korevaar DA, Cohen JF; STARD Group: STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015, 351:h5527
28. Tolles J, Lewis RJ: Time-to-event analysis. JAMA 2016, 315: 1046−1047
29. Stensrud MJ, Hernán MA: Why test for proportional hazards? JAMA 2020, 323:1401−1402
30. Cox DR: Regression models and life-tables. J R Stat Soc Ser B (Methodological) 1972, 34:187−220
31. Cheng DT, Mitchell TN, Zehir A, Shah RH, Benayed R, Syed A, Chandramohan R, Liu ZY, Won HH, Scott SN, Brannon AR, O'Reilly C, Sadowska J, Casanova J, Yannes A, Hechtman JF, Yao J, Song W, Ross DS, Oultache A, Dogan S, Borsu L, Hameed M, Nafa K, Arcila ME, Ladanyi M, Berger MF: Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): a hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. J Mol Diagn 2015, 17:251−264
32. Garcia EP, Minkovsky A, Jia Y, Ducar MD, Shivdasani P, Gong X, Ligon AH, Sholl LM, Kuo FC, MacConaill LE, Lindeman NI, Dong F: Validation of OncoPanel: a targeted next-generation sequencing assay for the detection of somatic variants in cancer. Arch Pathol Lab Med 2017, 141:751−758
33. Frampton GM, Fichtenholtz A, Otto GA, Wang K, Downing SR, He J, et al: Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. Nat Biotechnol 2013, 31:1023−1031
34. Beaubier N, Bontrager M, Huether R, Igartua C, Lau D, Tell R, Bobe AM, Bush S, Chang AL, Hoskinson DC, Khan AA, Kudalkar E, Leibowitz BD, Lozachmeur A, Michuda J, Parsons J, Perera JF, Salahudeen A, Shah KP, Taxter T, Zhu W, White KP: Integrated genomic profiling expands clinical options for patients with cancer. Nat Biotechnol 2019, 37:1351−1360
35. Mata DA, Benhamida JK, Lin AL, Vanderbilt CM, Yang S-R, Villafania LB, Ferguson DC, Jonsson P, Miller AM, Tabar V, Brennan CW, Moss NS, Sill M, Benayed R, Mellinghoff IK, Rosenblum MK, Arcila ME, Ladanyi M, Bale TA: Genetic and epigenetic landscape of IDH-wildtype glioblastomas with FGFR3-TACC3 fusions. Acta Neuropathol Commun 2020, 8:186
36. Bady P, Sciuscio D, Diserens A-C, Bloch J, van den Bent MJ, Marosi C, Dietrich P-Y, Weller M, Mariani L, Heppner FL, Mcdonald DR, Lacombe D, Stupp R, Delorenzi M, Hegi ME: MGMT methylation analysis of glioblastoma on the Infinium methylation BeadChip identifies two distinct CpG regions associated with gene silencing and outcome, yielding a prediction model for comparisons across datasets, tumor grades, and CIMP-status. Acta Neuropathol 2012, 124:547−560
37. Benhamida JK, Hechtman JF, Nafa K, Villafania L, Sadowska J, Wang J, Wong D, Zehir A, Zhang L, Bale T, Arcila ME, Ladanyi M:

Reliable clinical MLH1 promoter hypermethylation assessment using a high-throughput genome-wide methylation array platform. J Mol Diagn 2020, 22:368−375

38. Cao J, Zhang S: Multiple comparison procedures. JAMA 2014, 312: 543−544

39. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, Cerami E, Sander C, Schultz N: Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 2013, 6:pl1

40. Bahceci I, Dogrusoz U, La KC, Babur Ö, Gao J, Schultz N: PathwayMapper: a collaborative visual web editor for cancer pathways and genomic data. Bioinformatics 2017, 33: 2238−2240