Statistical considerations for outcomes in clinical research: A review of common data types and methodology

Matthew P Smeltzer and Meredith A Ray

Division of Epidemiology, Biostatistics, and Environmental Health, School of Public Health, The University of Memphis, Memphis, TN 38152, USA

Corresponding author: Matthew P Smeltzer. Email: msmltzer@memphis.edu

Impact Statement

Particularly in the clinical field, a larger variety of statistical analyses are conducted and results are utilized by a wide range of researchers, with some having more in-depth statistical training than others. Thus, we set out to summarize and outline appropriate statistical analyses for the most common data found in clinical research. We aimed to make this body of work comprehensive, yet brief and such that anyone working in clinical or public health research could gain a basic understanding of the different types of data and analyses.

Abstract

With the increasing number and variety of clinical trials and observational data analyses, producers and consumers of clinical research must have a working knowledge of an array of statistical methods. Our goal with this body of work is to highlight common types of data and analyses in clinical research. We provide a brief, yet comprehensive overview of common data types in clinical research and appropriate statistical methods for analyses. These include continuous data, binary data, count data, multinomial data, and time-to-event data. We include references for further studies and real-world examples of the application of these methods. In summary, we review common continuous and discrete data, summary statistics for said data, common hypothesis tests and appropriate statistical tests, and underlying assumption for the statistical tests. This information is summarized in tabular format, for additional accessibility.

Keywords: Statistics, epidemiology, analytic, biostatistics, outcomes, clinical

***Experimental Biology and Medicine* 2022; 247: 734–742. DOI: 10.1177/15353702221085710**

Introduction

Clinical research is vitally important for translating basic scientific discoveries into improved medical and public health practice through research involving human subjects.[1] The goal is to generate high-quality evidence to inform standards of care or practice. At its later stages, clinical research, in the form of clinical trials or observational studies, often focuses on comparing health outcomes between groups of persons who differ based on a treatment received or some other external exposure.[2]

The scientific method dictates that we test falsifiable hypotheses with quantitative data.[3] Evidence for or against a treatment or exposure must be evaluated statistically to determine if any observed differences are likely to represent true differences or are likely to have occurred by chance. Statistical methods are used to conduct hypothesis testing to this end.[4] In addition, statistical methods are employed to summarize the results of a study and to estimate the observed effect of the treatment or exposure on the outcome of interest.[4]

All clinical trials and many observational studies have a designated primary outcome of interest, which is the quantitative metric used to determine the effect of the treatment or exposure. The statistical properties, such as its probability distribution, of the outcome variable and quantifying changes in said variable due to the exposure are of primary importance in determining the choice of statistical methodology.[4,5] Here, we review some of the most common types of outcome variables in comparative research and common statistical methods used for analysis.

Approach

In this summary, we review standard statistical methodology used for data analysis in clinical research. We identify five common types of outcome data and provide an overview of the typical methods of analysis, effect estimates derived, and graphical presentation. We aim to provide a resource for the clinical researcher who is not a practicing statistician. Methods discussed can be reviewed in more detail in graduate-level textbooks on applied statistics, which are

ISSN 1535-3702

referenced throughout our summary. We also provide references for real-world clinical research projects that have employed each core method. In addition, the procedures available in standard statistical software for data analysis in each of these scenarios are provided in Supplemental Tables 1 and 2.

At the core, there are generally two categories of outcome data: discrete and continuous. By definition, discrete data, also called categorical data, are the data that have natural, mutually exclusive, non-overlapping groups.[6] Two examples would be severity, defined as mild, moderate, or severe, and intervention exposure groups, such as those receiving the intervention and those not receiving the intervention. Such categories may be ordinal (having an inherent order) or nominal (no inherent order) and can range from two groups or more.[6] The categories may represent qualitative groups (such as the previous examples) or quantitative data, that is, age groupings, such as 18–35, 36–55, 56 years and above.

Continuous data have more flexibility and can be defined as a variable that "can assume any values within a specified relevant interval of values."[6,7] More concrete examples include a person's age, height, weight, or blood pressure. While we may round for convenience, that is, round to the nearest integer (year) for age, there are no theoretical gaps between two continuous values. Addressing perhaps an obvious question, there are unique situations where data may skirt between discrete and continuous. For example, when does a quantitative ordinal discrete variable have enough categories to be considered a continuous variable? These are often a situation-by-situation basis and decided a priori before the onset of the study.

In addition to the type of data, the sample size may also influence the method used for calculating test statistics and *p*-values for statistical inference. When sample sizes are sufficiently large, we typically use a class of statistics called asymptotic statistics that rely on a result known as the central limit theorem.[8] These often rely on a Z statistic, chi-square statistic, or F statistic. When sample sizes are more limited, we typically use non-parametric or exact statistical methods that do not rely on these large sample assumptions. Most of the statistical methods that we review here rely on asymptotic statistics in their basic form, but often have an analogous method relying on exact and or non-parametric methods.[9] When a researcher encounters small sample sizes, it is important to consider these alternative methods.

In addition to identifying appropriate statistical methodology for testing hypotheses given the study's outcome data, there are a number of additional influences that should be considered, such as effect modification and confounding. Additional factors can alter the association of the exposure and outcome and thus are critical to consider when analyzing biological associations. Effect modification, by definition, occurs when a third factor alters the effect of the exposure on the outcome.[10] Specifically, the magnitude of this alteration changes across the values of this third factor. A separate phenomenon, known as confounding, occurs when an imbalance in the distribution of a third factor in the data distorts the observed effect (association) of the exposure on the outcome.[10] To meet the criteria of a confounder, this third factor must be associated with the exposure and with the outcome but not in the casual pathway. If all scenarios above occur, this third factor is a confounder and introduces bias when not properly controlled. While effect modification is a biologic phenomenon in which an exposure impacts the outcome differently for different groups of individuals, confounding is a phenomenon caused by the imbalance of the data itself and may not have biologic significance.

An important consideration is that the effect-modifying or confounding factor is not in the casual pathway from the exposure to outcome. The casual pathway is the primary biological pathway in which the exposure influences the outcome. For example, if Variable A causes Z, and Z causes Y, then Variable Z is on the causal pathway from A to Y. In this case, controlling for Z as either a confounder or effect modifier while estimating the effect of A on Y will induce bias in the estimate. Investigators should also avoid controlling for common effects of A and Y, which can induce "collider bias." We will discuss how to assess for effect modification and confounding later.

Methods for common variable types

Continuous data

Continuous data, as described above, are quantitative data with no theoretical gaps between values, where the range (minimum and maximum) of values is dependent on what is being measured. For example, the natural range for age is (0, ~100) while the natural range for temperature measured in degrees Fahrenheit is (−459.67, 134). These types of data are often summarized with a central measurement and a spread measurement.[7] The most common central measurements are the mean or median and represent the "center" of the observed data. The spread measurement aims to quantify how much variation is in the data or how much of the data deviates from the central measurement. Thus, if a mean is presented as the central metric, the variance or standard deviation is typically presented as the spread measurement. If the median is presented as the central metric, the interquartile range (IQR: 25th and 75th percentile) and range (minimum and maximum) are reported as the spread measurement. Understandably, the next question is: which metric to use and when?

This leads to the topic of data distributions. If our continuous data follow what we call a normal probability distribution, this is a symmetric distribution around the mean and therefore, our mean and median will be approximately the same value.[7] While it is statistically appropriate to report either the mean (with variance or standard deviation) or the median (with IQR and range), if our data follow a normal distribution, the most common practice is to report the mean. If the data are skewed and do not follow a normal distribution, it is appropriate to report the median. If the data are skewed, the mean is pulled toward the more extreme values and no longer a true central measurement, while the median is not influenced by skewness.[7]

A normal distribution is a statistical probability distribution, defined by a mean and variance, that illustrates the probability of observing a specific value or values from the data. It has convenient statistical properties, such as a

Table 1. Summary of continuous data analyses and assumptions (all observations are independent).

Type of outcome variable	Outcome statistical distribution	Theoretical hypotheses*	Assumptions	Commonly used point estimate	Commonly used Effect estimate[13–15]	Common statistical methods
One variable	Normal	$H_0 : \mu = \delta_0$ $H_1 : \mu \neq \delta_0$	Normality	Mean	Mean	*T*-test
	No assumption	$H_0 : M = \delta_0$ $H_1 : M \neq \delta_0$	None	Median	Median	Sign test or signed-rank test
Two variables	Normal	$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$	1. Normality 2. Two groups are independent 3. Group variances are equal	Mean	Difference of means or Cohen's d	*T*-test
	No Assumption	$H_0 : M_1 = M_2$ $H_1 : M_1 \neq M_2$	1. Two groups are independent 2. Both groups have same distribution shape	Median	U statistic	Mann–Whitney U test**
Three or more variables (*k* groups)	Normal	$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ $H_1 :$ *at least one mean is different*	1. Normality 2. All groups are independent 3. Group variances are equal	Mean	Cohen's f	ANOVA
	No Assumption	$H_0 : M_1 = M_2 = \ldots = M_k$ $H_1 :$ *atleast one median is different*	1. All groups are independent	Median	ε^2	Kruskal–Wallis
Association analyses: modeling outcome as a function of one or more explanatory variables						
One continuous variable	Normal	$H_0 : \beta_1 = \ldots = \beta_k = 0$ $H_1 :$ *at least one* β *is not* 0 or $H_0 : \beta_i = 0$ $H_1 : \beta_i \neq 0,$ *for* $i = 1,\ldots,k$ *explanatory variables*	1. Linear association between explanatory variables and outcome 2. Independent explanatory variables (if more than one) 3. Normally distributed error terms 4. Equal variances	None	Cohen's f or R^2	Linear regression (overall F-test or partial *t*-tests)

ANOVA: Analysis of variance.

*μ = mean, δ_0 = value of interest, M = median, β = regression coefficients.

**has multiple names.

Table 2. Summary of discrete data analyses and assumptions (all observations are independent).

Type of outcome variable	Outcome statistical distribution	Theoretical hypotheses*	Assumptions	Commonly used point estimate	Commonly Used Effect estimate[13–15]	Common statistical methods
Discrete						
One binary variable	Binomial	$H_0: \pi = \delta_0$ $H_1: \pi \neq \delta_0$	One binary variable.	Proportion	Proportion	Z-test or binominal exact test
Two binary variables	Binomial	$H_0: \pi_1 = \pi_2$ $H_1: \pi_1 \neq \pi_2$	1. One binary metric measured on two different samples. 2. Two samples are independent.	Proportions	Difference in proportions or Cohen's h	Z-test
		$H_0: OR = 1$ $H_1: OR \neq 1$	1. Two binary variables measured on same sample. 2. One variable measuring outcome. 3. One variable measuring exposure.	Odds	Odds ratio	Logistic regression
		$H_0: RR = 1$ $H_1: RR \neq 1$	1. Two binary variables measured on same sample. 2. One variable measuring outcome. 3. One variable measuring exposure.	Risk	Risk ratio	Logistic, Poisson, or negative binomial regression
Two discrete variables	No Assumption	H_0: *There is no association between the two variables* H_1: *There is an association between the two variables*	1. Two variables measured on the same sample. 2. Each variable is measuring a different metric.	None	Cramer's V or Phi	Chi-squared test, Fisher's exact test (small sample sizes), or logistic regression
Association analyses: modeling outcome as a function of one or more explanatory variables						
One binary variable	Binomial	$H_0: \beta_1 = \ldots = \beta_k = 0$ H_1: *at least one* β *is not* 0 or $H_0: \beta_i = 0$	1. Outcome variable is binary 2. Explanatory variables are independent 3. Explanatory variables are linearly associated with the log odds.	Odds	Odds ratio	Logistic regression
One discrete variable with > 2 levels	Multinomial (ordered or unordered)	$H_1: \beta_i \neq 0$ or $H_0: OR_i = 1$ $H_1: OR_i \neq 1$ *for* $i = 1, \ldots, k$ *explanatory variables*	If outcome data are nominal, the assumptions are the same as binomial logistic regression. If outcome data are ordinal, the proportional odds assumption must be met in addition to binomial logistic regression assumptions.	Odds	Odds ratio	Multinomial logistic regression: generalized logit link for unordered and cumulative logit link for ordered

(Continued)

Table 2. (Continued)

Type of outcome variable	Outcome statistical distribution	Theoretical hypotheses*	Assumptions	Commonly used point estimate	Commonly Used Effect estimate[13–15]	Common statistical methods
Counts and events per follow-up	Poisson or negative binomial	$H_0: \beta_1 = \ldots = \beta_k = 0$ $H_1:$ *at least one* β *is not* 0 or $H_0: \beta_i = 0$ $H_1: \beta_i \neq 0$ or $H_0: IRR_i = 1$ $H_1: IRR_i \neq 1$ *for* $i = 1, \ldots, k$ exp*lanatory variables*	1. Outcome variable is positive integer counts following a Poisson or negative binomial distribution	Incidence rate	Incidence rate ratio	Poisson or negative binomial regression
Time-to-event	No Distribution Assumed	$H_0: S(t)_1 = S(t)_2 = \ldots = S(t)_k$ $H_1:$ *at least one survival curve is different*	1. Single discrete exploratory variable (with *k* categories) 2. Censoring is not related to explanatory variables	5-year survival	Difference in 5-year survival	Kaplan–Meier (Log-rank test)
		$H_0: \beta_1 = \ldots = \beta_k = 0$ $H_1:$ *at least one* β *is not* 0 or $H_0: \beta_i = 0$ $H_1: \beta_i \neq 0$ or $H_0: HR_i = 1$ $H_1: HR_i \neq 1$ *for* $i = 1, \ldots, k$ *explanatory variables*	1. Hazard remains constant over time (hazards are proportional assumption). 2. Explanatory variables are independent. 3. Explanatory variables are linearly associated with the log hazard.	None	Hazard ratio	Cox proportional hazards model

*π = proportion, δ_0 = value of interest, OR = odds ratio, RR = risk ratio, β = regression coefficients, IRR = incident rate ratio, $S(t)$ = survival function/curve, HR = hazard ratio.

pre-specified probability density function and cumulative density function, which are the functions that calculate said probabilities.[7] In addition to the normal distribution, other distributions exist for continuous data and discrete data. Other continuous distributions include, but not limited to exponential, chi-square, F, T, gamma, and beta distributions.[8] Discrete distributions include, but not limited to Bernoulli, binomial, Poisson, negative binomial, and hypergeometric.[8] Each distribution is defined by one or more parameters which control the average, standard deviation, and other aspects of the distribution. If the data follow one of these known distributions, calculating the likelihoods of occurrence, such as for hypothesis testing, becomes straightforward.

How we determine if data follow one of these distributions vary for each type of distribution. For the scope of this body of work, we will only cover how to assess if a continuous variable follows a normal distribution. There are three ways in which one can assess normality, each has its strength and weakness and, therefore, encourage the consideration of all three approaches. Normality can be assessed visually with quantile–quantile (QQ) plots, visually with histograms, or by statistical test (Shapiro–Wilk test, Kolmogorov–Smirnov test, Cramer–von Mises test, or Anderson–Darling test).[11,12] Other tests exist but these are the most commonly available in statistical software. The normality tests tend to be very strict, and the smallest deviations will lead to non-normal conclusion.[11,12] The visual assessments, such as the QQ plots and histograms, are more subjective to the researcher's judgment, hence useful to consider visual and statistical approaches.

When our outcome variable is normally distributed, there are several factors that must be considered for selecting the appropriate statistical method to test the hypothesis, such as number of samples, independence, and so on. These analyses have been summarized in Table 1. Note this table is not comprehensive but a generalized summary of common analyses and assumptions. When the continuous outcome violates normality, or the sample size is small, non-parametric approaches can instead be used. Non-parametric approaches are analyses that do not make any assumptions about the type of distribution; they can analyze normal and non-normally distributed data. However, if data are normal, parametric approaches are more appropriate to implement.

If our aim is to quantify the association between an outcome and exposure, we can apply linear regression (assuming all assumptions are met, see Table 1). As outlined earlier, we need to consider possible effect modifiers and confounders. To assess for effect modification, we can do so by introducing an interaction term in the model. As a simple example, the model would contain the exposure variable, the possible effect modifier, and a multiplication term between the exposure and possible effect modifier (termed the interaction term). If the interaction term is statistically significant, we would conclude effect modification is present. If a variable is not an effect modifier, consideration for confounding is then checked. There exist different approaches for assessing confounding but the most widely used is the 10% rule. This rule states that a variable is a confounder if the regression coefficient for the exposure variable changes by more than 10% with the inclusion of the possible confounder in the model. A nice example of this can be seen in Ray *et al.* (2020).[16]

Counts and rates

Count data are the number of times a particular event occurs for each individual, taking non-negative integer values. In biomedical science, we most often look at count data over a period of time, creating an event rate (event count / period of time). The simplest analysis of these data involves calculating events per patient-year of follow-up. When conducting patient-year analyses in large populations, it is often acceptable to look at this statistic in aggregate (sum of total events in the population / sum of total patient-years at risk in the population). Confidence intervals can be calculated by assuming a Poisson distribution.

Statistical modeling of count data or event rates is common with a Poisson model. These models can adjust for confounding by other variables and incorporate interaction terms for effect modification. When a binary treatment variable is used with event rate as the outcome, incidence rate ratios (with confidence intervals) can be estimated from these models. The model can be extended to a zero inflated Poisson (ZIP) model or a negative binomial model when the standard Poisson model does not fit the data well. Population level analyses often look at disease incidence rates and ratios using these methods.[17,18] Recently, this type of statistic modeling is at the core of statistical methods used to calculate vaccine efficacy against COVID-19 in a highly impactful randomized trial.[19]

Binary data

Arguably, the simplest form of an outcome variable in clinical research is the binary variable for which every observation is classified in one of two groups (disease versus no disease, response versus no response, etc.).[20] We typically assume a binomial statistical distribution for this type of data. When the treatment variable is also binary, results can be analyzed by the simple analysis of the classic 2 × 2 table. From this table, we can estimate the proportion of responses, odds of response, or risk of response/disease within each treatment group. We then compare these estimates between treatment groups using differences or ratio measures. These include the difference in proportions, risk difference, odds ratios, and risk ratios (relative risk). Hypothesis testing around these estimates may utilize the chi-square test to assess the general association between the two variables, large sample asymptotic tests relying on normality under the central limit theorem, or exact tests that do not assume a specific statistical distribution.

Statistical models for binary outcomes can be constructed using logistic regression. In this way, the effect estimates (typically the odds ratio) can be adjusted for confounding by measured variables. These models typically rely on asymptotic normality for hypothesis testing but exact statistics are also available. The models can also assess effect modification through statistical interaction terms. An example of the classical 2 × 2 table can be referenced in Khan *et al.*[21] A typical application of logistic regression can be seen in Ray *et al.*[22] We have summarized methods for categorical data in Table 2.

Multinomial data

Multinomial data are a natural extension of binary data such that it is a discrete variable with more than two levels.

It follows that the extensions of logistic regression can be applied to estimate effects and adjust for effect modification and confounding. However, multinomial data can be nominal or ordinal. For nominal data, the order is of no importance and, therefore, the models use a generalized logit link.[23] This will select one category as a referent category and then perform a set of logistic regression models, each comparing one non-referent level to this referent level. For example, in Kane *et al.*,[24] they applied a multinomial logistic regression to model type of treatment (five categories) as a function of education level and other covariates. They select watchful waiting as the referent treatment. The analysis thus had four logistic regressions to report, respective of each of the other treatment categories compared to watchful waiting.

If the multinomial data are ordinal, we use a cumulative logit link in the regression model. This link will model the categories cumulatively and sequentially.[23] For example, suppose our outcome has three levels, 1, 2, and 3 and are representative of the number of treatments. Cumulative logit will conduct two logistic regressions: first, Modeling Category 1 versus Categories 2 and 3 (combined) and then Categories 1 and 2 (combined) versus Category 3. Because of the combining of categories, this assumes that the odds are proportional across categories. Thus, this assumption must be checked and satisfied before applying this model. Depending on the outcome, only one of the logistic models may be needed, such as in Bostwick *et al.*,[25] where their outcome was palliative performance status (low, moderate, and high) and the effects of cancer/non-cancer status. Here, they only reported high-performance status versus moderate and low combined as their outcome.

Time-to-event

Time-to-event data, often called survival data, compare the time from a baseline point to the potential occurrence of an outcome between groups.[26] These data are unique as a statistical outcome because they involve a binary component (event occurred or event did not occur) and the time to event occurrence or last follow-up. Both the occurrence of event and the time it took to occur are of interest. These outcomes are most frequently analyzed with two common statistical methodologies, the Kaplan–Meier method and the Cox proportional hazards model.[26]

The Kaplan–Meier method allows for the estimation of a survival distribution of observed data in the presence of censored observations and does not assume any statistical distribution for the data.[26,27] In this way, knowledge that an individual did not experience an event up to a certain time point, but is still at risk, is incorporated into the estimates. For example, knowing an individual survived 2 months after a therapy and was censored is less information than knowing an individual survived 2 years after a therapy and was censored. The method assumes that the occurrence of censoring is not associated with the exposure variable. In addition to estimating the entire curve over time, the Kaplan–Meier plot allows for the estimation of the survival probability to a certain point in time, such as "5-year" survival. Survival curves are typically estimated for each group of interest (if exposure is discrete), shown together on a plot. The log-rank test is often used to test for a statistically significant difference in two or more survival curves.[26] An analogous method, known as Cumulative Incidence, takes a similar approach to the non-parametric Kaplan–Meier method, but starts from zero and counts events as they occur, with estimates increasing with time (rather than decreasing).[26] Cumulative Incidence analyses can also be adjusted for competing risks, which occur when subjects experience a different event during the follow-up time that precludes them from experiencing the event of primary interest. In the presence of competing risks, Cumulative Incidence curves can be compared using Gray's test.[26]

Time-to-event data can also be analyzed using statistical models. The most common statistical model is the Cox proportional hazards model.[28] From this model, we can estimate hazard ratios with confidence intervals for comparing the risk of the event occurring between two groups.[26] Multiple variable models can be fit to incorporate interaction terms or can be adjust for confounding (the 10% rule can be applied to the hazard ratio estimate). Although the Cox model does not assume a statistical distribution for the outcome variable, it does assume that the ratio of effect between two treatment groups is constant across time (i.e., proportional hazards). Therefore, one hazard ratio estimate applies to all time points in the study. Extensions of this model are available to allow for more flexibility, with additional complexity in interpretation. Examples of standard applications of the Kaplan–Meier method and Cox proportional hazards models can be seen in recent papers by Mok *et al.*[29] and Aparicio *et al.*[30]

Generalized linear models

With the exception of time-to-event data, all of the statistical modeling techniques described above can be classified as some form of generalized linear model (GLM).[20] Modern statistical methods utilize GLMs as a broader class of statistical model. In the GLM, the outcome variable can take on different forms (continuous, categorical, multinomial, count, etc) and it is mathematically transformed using a link function. In fact, the statistical modeling methods we have discussed here are each a special case of a GLM. The GLM can accommodate multiple covariates that could be either continuous or categorical. The GLM framework is often a useful tool for understanding the interconnectedness of common statistical methods. For the interested reader, an elegant description of the most common GLMs and how they interrelate is given in Chapter 5 of *Categorical Data Analysis* by Alan Agresti.[20]

Concerns of bias and validity

While statistical significance is necessary to demonstrate that an observed result is not likely to have occurred by chance alone, it is not sufficient to insure a valid result. Bias can arise in clinical research from many causes, including misclassification of the exposure, misclassification of the outcome, confounding, missing data, and selection of the study cohort.[10,31] Care should be taken at the study design phase to reduce potential bias as much as possible. To this end, application of proper research methodology is essential. Confounding can sometimes be corrected through statistical adjustment after collection of the data, if the confounding factor is properly measured in the study.[10,31] All of these issues are outside the scope of basic statistics and this current summary. However, good clinical research studies should consider both statistical methodology and potential threats to validity from bias.[10,31]

Discussion

In this review, we have discussed five of the most common types of outcome data in clinical studies, including continuous, count, binary, multinomial, and time-to-event data. Each data type requires specific statistical methodology, specific assumptions, and consideration of other important factors in data analysis. However, most fall within the overarching GLM framework. In addition, the study design is an important factor in the selection of the appropriate method. Statistical methods can be applied for effect estimation, hypothesis testing, and confidence interval estimation. All of the methods discussed here can be applied using commonly available statistical analysis software without excessive customized programming.

In addition to the common types of data discussed here, other statistical methods are sometimes necessary. We have not discussed in detail situations where data are correlated or clustered. These scenarios typically violate the independence assumption required by many methods. Common subsets of these include longitudinal analyses with multiple observations collected across time and time series data which also require specialized techniques. We have also not covered situations where outcome data are multidimensional, such as the case for research in genetics. The analysis of large amounts of genetic information often relies on the basic methods discussed here, but special considerations and adapted methodology are needed to account for the large numbers of hypothesis tests conducted. One consideration is multiple comparisons. When a single sample is tested more than one time, this increases the chance of making either type I or II error.[32] This means we incorrectly reject or fail to reject the null hypothesis given the truth at the population level. Because of this increased likelihood of error, the significance level must be adjusted. These types of adjustments are not discussed here. Moreover, this overview is not comprehensive, and many additional statistical methodologies are available for specific situations.

In this work, we have focused our discussion on statistical analysis. Another key element in clinical research is a priori statistical design of trials. Appropriate selection of the trial design, including both epidemiologic and statistical design, allows data to be collected in a way that valid statistical comparisons can be made. Power and sample size calculations are key design elements that rely on many of the statistical principals discussed above. Investigators are encouraged to work with experienced statisticians early in the trial design phase, to ensure appropriate statistical considerations are made.

In summary, statistical methods play a critical role in clinical research. A vast array of statistical methods are currently available to handle a breath of data scenarios. Proper application of these techniques requires intimate knowledge of the study design and data collected. A working knowledge of common statistical methodologies and their similarities and differences is vital for producers and consumers of clinical research.

AUTHORS' CONTRIBUTIONS

All authors participated in the design, interpretation of the studies, writing and review of the manuscript.

DECLARATION OF CONFLICTING INTERESTS

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

FUNDING

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID ID

Matthew P Smeltzer https://orcid.org/0000-0003-1366-9267

SUPPLEMENTAL MATERIAL

Supplemental material for this article is available online.

REFERENCES

1. Habets MG, van Delden JJ, Bredenoord AL. The social value of clinical research. *BMC Med Ethics* 2014;**15**:17
2. FDA. What are the different types of clinical research? 2018, https://www.fda.gov/patients/clinical-trials-what-patients-need-know/what-are-different-types-clinical-research
3. Popper K. *The logic of scientific discovery*. London: Hutchinson and Co, 1959
4. Armitage P, Berry G, Matthews JNS. *Statistical methods in medical research*. Chichester: John Wiley & Sons, 2008
5. Valentinuzzi ME, Friedman LM, Furberg CD, DeMets DL. *Fundamentals of clinical trials*. 3rd ed. Cham: Springer, 2004
6. Moore DS, Dawson LG, McCabe GP. *Introduction to the practice of statistics excel manual with macros*. London: Macmillan, 2005
7. Daniel WW, Cross CL. *Biostatistics: A foundation for analysis in the health sciences*. Chichester: Wiley, 2018
8. Casella G, Berger RL. *Statistical inference*. Belmont, CA: Cengage Learning, 2021
9. Hollander M, Wolfe DA, Chicken E. *Nonparametric statistical methods*. Chichester: John Wiley & Sons, 2013
10. Lash TL, VanderWeele TJ, Haneuse S, Rothman KJ. *Modern epidemiology*, 4th ed. New York, NY: Wolters Kluwer, 2021
11. Thode HC. *Testing for normality*. Boca Raton, FL: CRC press, 2002
12. Razali NM, Wah YB. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *J Stat Model Anal* 2011; **2**:21–33
13. Serdar CC, Cihan M, Yücel D, Serdar MA. Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies. *Biochem Med* 2021;**31**:27–53
14. Dey R, Mulekar MS. Effect size as a measure of difference between two populations. In: Alhajj R and Rokne J (eds) *Encyclopedia of social network analysis and mining*. New York: Springer, 2018. pp.715–26
15. Conroy RM. What hypotheses do "nonparametric" two-group tests actually test? *Stata J* 2012;**12**:182–90
16. Ray MA, Smeltzer MP, Faris NR, Osarogiagbon RU. Survival after mediastinal node dissection, systematic sampling, or neither for early stage NSCLC. *J Thorac Oncol* 2020;**15**:1670–81
17. Koton S, Sang Y, Schneider AL, Rosamond WD, Gottesman RF, Coresh J. Trends in stroke incidence rates in older us adults: an update from the atherosclerosis risk in communities (ARIC) cohort study. *JAMA Neurol* 2020;**77**:109–13

18. Lei J, Ploner A, Elfström KM, Wang J, Roth A, Fang F, Sundström K, Dillner J, Sparén P. HPV vaccination and the risk of invasive cervical cancer. *N Engl J Med* 2020;**383**:1340–8
19. Polack FP, Thomas SJ, Kitchin N, Absalon J, Gurtman A, Lockhart S, Perez JL, Marc GP, Moreira ED, Zerbini C. Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. *N Engl J Med* 2020;**383**:2603–15
20. Agresti A. *Categorical data analysis*. Chichester: John Wiley & Sons, 2003
21. Khan NE, De Souza A, Mister R, Flather M, Clague J, Davies S, Collins P, Wang D, Sigwart U, Pepper J. A randomized comparison of off-pump and on-pump multivessel coronary-artery bypass surgery. *N Engl J Med* 2004;**350**:21–8
22. Ray MA, Faris NR, Derrick A, Smeltzer MP, Osarogiagbon RU. Rurality, stage-stratified use of treatment modalities, and survival of non-small cell lung cancer. *Chest* 2020;**158**:787–96
23. Harrell FE. Regression modeling strategies. *Bios* 2017;**330**:14
24. Kane CJ, Lubeck DP, Knight SJ, Spitalny M, Downs TM, Grossfeld GD, Pasta DJ, Mehta SS, Carroll PR. Impact of patient educational level on treatment for patients with prostate cancer: data from CaPSURE. *Urology* 2003;**62**:1035–9
25. Bostwick D, Wolf S, Samsa G, Bull J, Taylor DH Jr, Johnson KS, Kamal AH. Comparing the palliative care needs of those with cancer to those with common non-cancer serious illness. *J Pain Symptom Manage* 2017;**53**:1079–84
26. Kleinbaum D, Klein M. *Survival analysis: A self-learning text, 2005*. New York: Springer-Verlag, 2011
27. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc* 1958;**53**:457–81
28. Cox DR. Regression models and life-tables. *J Roy Stat Soc: Ser B* 1972;**34**:187–202
29. Mok TS, Wu Y-L, Ahn M-J, Garassino MC, Kim HR, Ramalingam SS, Shepherd FA, He Y, Akamatsu H, Theelen WS. Osimertinib or platinum–pemetrexed in EGFR T790M–positive lung cancer. *N Engl J Med* 2017;**376**:629–40
30. Aparicio HJ, Himali JJ, Beiser AS, Davis Plourde KL, Vasan RS, Kase CS, Wolf PA, Seshadri S. Overweight, obesity, and survival after stroke in the Framingham Heart Study. *J Am Heart Assoc* 2017;**6**:e004721
31. Weisberg HI. *Bias and causation: Models and judgment for valid comparisons*. Chichester: John Wiley & Sons, 2011
32. Chen S-Y, Feng Z, Yi X. A general introduction to adjustment for multiple comparisons. *J Thorac Dis* 2017;**9**:1725–9