Chapter 15
The Application of Biostatistics to Your Surgical Practice

Vlad V. Simianu, Mark Pedersen, Rebecca P. Petersen, and Anjali S. Kumar

Introduction

As academic centers recruit and hire junior faculty to fill the large shoes of senior surgeons who are retiring or are promoted to new positions, the number of new responsibilities that are foreign to a recent graduate of a surgical training program can be overwhelming. In addition to being the

V.V. Simianu, M.D., M.P.H.
Surgical Outcomes Research Center, University of Washington, Seattle, WA, USA

Department of Surgery, University of Washington, Seattle, WA, USA

M. Pedersen, M.D.
Department of Surgery, University of Iowa Hospitals and Clinics, Iowa City, IA, USA

R.P. Petersen, M.D., M.S.
Department of Surgery, University of Washington, Seattle, WA, USA

A.S. Kumar, M.D., M.P.H. (✉)
Department of Surgery, Virginia Mason Medical Center, Seattle, WA, USA
e-mail: askumarmd@gmail.com

D.B. Renton et al. (eds.), *The SAGES Manual Transitioning to Practice*, DOI 10.1007/978-3-319-51397-3_15

primary point-person of complex patient panels, young surgeons also juggle medical record upkeep, billing, and practice promotion. Those who land jobs with academic appointments may be asked to perform and publish research. Productivity in these arenas may influence the new faculty member's ability to rise in rank over the years, thus affecting salary.

Even when/if the position is exempt from performing research, teaching faculty will be asked to participate and/or lead journal review conferences for the department or section. An intelligent and insightful review of the literature selected can make an indelible impression on a colleague in the community. Arrive prepared to discuss the studies that have been selected by using this chapter and its pearls (Fig. 15.1) as your primer to the daunting, but decipherable, world of biostatistics.

What Is the Study's Purpose?

A study's purpose drives the selection of data sources, outcomes of interest, study design, and analytic plan. In general, the purpose of a study falls into two categories: hypothesis-generating or hypothesis-testing. Hypothesis-generating (sometimes called descriptive) studies aim to identify possible associations and motivate future investigations. Hypothesis-testing studies should make clear whether the hypothesis concerns superiority, inferiority, or equivalence (non-inferiority), and every attempt to exclude the influence of chance and bias in evaluating the hypothesis.

To succinctly summarize the research question that a study aims to address, we recommend that the discussant uses the "PICOT framework:" **P**opulation, **I**ntervention (i.e., independent variables, exposure, and covariates), the **C**omparator group (if applicable), **O**utcome (i.e., dependent variable or endpoint), and **T**ime frame of outcomes assessment [1–3]. Approaching research studies through a PICOT lens can guide the reader/reviewer systematically through the pertinent considerations when appraising the work.

- The research question can be succinctly summarized using the "PICOT" framework (Population, Intervention, Comparator group, Outcome, and Time frame)
- Recognizing of the minimal-clinically important difference (or MCID) for a particular patient-reported outcome (PRO) is helpful to help the reader distinguish between results that are significant statistically and those which are perceived as clinically significant to patients.
- ITT (intention to treat) analysis is essential because it provides information about how the intervention compares at the moment the decision is being made and is particularly useful at counseling patients.
- If the data are skewed (not normally distributed) the mean will be a biased estimator of the central tendency. In these cases, the median provides a better estimate.
- As a general rule, the larger the difference being compared and the larger the sample size for a given comparison, the lower the *p* value, and the less likely that the finding is the result of chance alone.
- When multiple comparisons are necessary, corrections (e.g., Bonferonni correction) to appropriately lower the p-value can be made in an attempt to safeguard against Type I errors (a false-positive finding).
- *Type II error* (a false-negative finding) most commonly occurs when a study has insufficient power (insufficient size) to detect true differences in outcomes between groups.
- When the summary measure is the absolute difference or relative risk, a CI inclusive of 0 indicates no statistically significant difference. If the summary measure is an odds ratio, a CI inclusive of 1.0 indicates no statistical difference in outcomes.
- When the continuous variable is not normally distributed, an alternative to the t-test, such as the Wilcoxon rank-sum test, may be more appropriate
- Fisher's exact test is more appropriate for such comparisons when the sample size is small (<100).
- As a rule of thumb, a minimum of 10 events (and equivalent number of nonevents) per variable are required for logistic regression (binary outcome) and 10 to 15 observations per variable for linear regression (continuous outcome).
- Odds ratio will overestimate the probability if the outcome occurs frequently (>10%) in the population.

Fig. 15.1 Pearls: tips and tricks for deciphering statistical implications of studies

Is the Right Data Being Used?

Many sources of information exist to conduct clinical research, and the selection of a data source is driven by a balance between the study purpose, resources (i.e., money), and feasibility (i.e., acceptance, ethics, and time). Table 15.1—Data sources provides a synopsis of commonly used data sources. The strengths and limitations of each are highlighted. For example, many datasets rely completely on administrative data (i.e., Medicare claims). These datasets are readily available to researchers and relatively inexpensive to obtain and analyze. However, they only reliably include metrics related to the billable aspects of care.

What Is the Measured Outcome?

Outcomes assessment cannot determine which intervention is better for the patient, but it can inform patients and providers about differences between competing diagnostic or therapeutic options. It is therefore important to determine which outcomes were assessed in a study, from what perspective, and whether these were consistent with the study's purpose. Outcomes may be subjective (e.g., patient satisfaction) or objective (e.g., death). There are categories of outcomes that the study designer or evaluator should be familiar with: (1) clinical outcomes, (2) patient-reported outcomes (PROs), (3) financial outcomes.

Clinical outcomes are well-defined, validated, and relatively easy to measure. Those related to in-hospital safety (safety outcomes) only require a short follow-up period. Operative mortality and postoperative complications (morbidity) are the most commonly measured safety endpoints. While, fortunately, mortality and complications tend to be infrequent events, the statistical implications are that safety studies often need to be quite large in sample size to be able to identify differences. One approach to address the costs of large samples or long lag times of development of infrequent

TABLE 15.1 Common data sources and their unique advantages and disadvantages (adapted, with permission, from Simianu et al. [17])

Data source	Advantages	Disadvantages	Example
Medical records	Easy to obtain Useful for hypothesis generation	Missing data Time-consuming Inability to measure certain information (e.g., intent) Limited scientific value	Case reports Case series
Patient-Reported Outcomes (PRO)	Unique data on symptoms, function, and health status Global (multidimensional) or Specific (unidimensional)	Time-consuming: interviews or questionnaires Unique instruments can have validity issues when population broadened Change/effect can be difficult to interpret	SF-36 Health Survey Patient-Reported Outcomes Measurement Information System (PROMIS)
Registry	Often contains clinical data Population-based real-world data not restricted to tertiary or referral centers	Built for limited reasons, so has restricted data Often has missing data because information captured from usual care rather than research visits Often include only cross-sectional data and need linkage to other data sources for follow-up	SEER National Cancer Database Device registries (Transcatheter Aortic Valve Replacement, TAVR)

(continued)

Table 15.1 (continued)

Data source	Advantages	Disadvantages	Example
National surveys	National sample Some longitudinal diagnoses and health care claims data	May over-represent certain racial groups in survey sample	Medical expenditure panel survey
Quality Improvement and Surveillance Project	Prospectively collected data Rich in clinical, laboratory, and demographic patient data	Overrepresentation of tertiary or referral centers Only a random sample of patients, not comprehensive	National Surgical Quality Improvement Project Society of Thoracic Surgeons database
Administrative	Large numbers Real-world data Often generalizable Easy to obtain Affordable	Limited clinical variables Data collected for billing, not research	Medicare State discharge data
Linked datasets	Richer source of data than either registry or administrative alone Allows longitudinal assessment of episodes of care	Missing data Inability to capture intent of therapy	SEER-Medicare

SEER surveillance, epidemiology, and end results program

clinical events is to report *surrogate endpoints*. Surrogate endpoints are intermediate outcomes that might serve as a surrogate for the actual clinical effect. However, the reader must consider whether the selected outcome is a meaningful clinical endpoint or simply a more easily measured surrogate [4]. Alternatively, when events are rare or there is no single optimal outcome, studies may report *composite endpoints*. For composite endpoints to be meaningful, however, they should be of similar importance and frequency. Imbalance in the components will not allow reviewers to judge which individual outcome contributed most to the composite endpoint.

Patient-reported outcomes (PROs) measure experiences or events that are reported by the patient. Sometimes, PROs are regarded as subjective outcomes because the response cannot be verified by a provider or researcher. Examples of common PRO concepts are health-related quality of life (HRQOL), satisfaction with care, functional status, well-being, and health status. Discrete concepts (PRO domains), include physical (e.g., pain), psychological (e.g., depression), and social functioning (e.g., the ability to carry out activities of daily living). Researchers are advised to use existing instruments to measure PROs (rather than creating their own) because the appropriate development of a questionnaire requires significant time, resources, testing, and validation before application. Recognition of the minimal-clinically important difference for a particular PRO distinguishes between results that are significant statistically and those which are perceived as clinically significant to patients [5].

Increasingly, *financial outcomes* are being added to contemporary studies. In these scenarios, it is important to note the difference between charges (the amount of money requested for health services and supplies) and costs (the actual amount of money spent). In addition, it is important to recognize that handling cost data requires special statistical approaches because costs are highly skewed (a few patients experience disproportionately higher costs than the majority) and exist as point masses (where many patients incur no costs). As an alternative to cost data, some authors report

resource utilization which can range from pre-hospital resources (such as clinic visits and preoperative tests) to hospital resources (length of stay, readmissions, pharmacy services) to post-hospital care (skilled nursing facilities and home care). Similar to cost data, resources utilization data suffers from skewing and clustering. Three common approaches to cost-outcomes are cost-benefit, cost-utility, and cost-effectiveness analyses.

What Is the Hypothesis Being Tested?

Hypothesis testing is used to determine whether observed differences between two or more groups are true findings or are attributable to chance alone. Prior to testing a hypothesis, it is important to first define the null and alternative hypotheses. A *null hypothesis* is the principle that there is no difference among groups. The *alternative hypothesis* is the idea that there is a difference among two groups. A researcher needs to know with an acceptable level of accuracy whether an outcome is occurring due to the alternative hypothesis being correct or by chance alone. The *p*-value is a statistical summary measure for hypothesis testing and is interpreted as the probability that the observed difference in outcomes between groups is the result of chance (i.e., the difference is not actually based on the effect of the intervention). A significant level of 5% ($p = 0.05$) is widely accepted in medical literature to indicate a statistically significant finding. This threshold is rather arbitrary and for some measures a lower (large databases where false positive are to be avoided) or higher level (when a higher noise-to-signal ratio is acceptable as in safety evaluations) may be appropriate. As a general rule, the larger the difference being compared and the larger the sample size for a given comparison, the lower the *p*-value, and the less likely that the finding is the result of chance alone.

There are two types of errors which can occur with any hypothesis testing. Understanding how to address them is pertinent to the study purpose, design, and analytic plan. A *type 1 error* or *false positive* occurs when one observes a

difference in outcomes when one does not actually exist. In this case, the null hypothesis is incorrectly rejected. For example, if a threshold of 5% ($p < 0.05$) were considered statistically significant, 5 of 100 statistical tests could potentially demonstrate a statistically significant finding that is attributable to chance alone. If one repeats a comparative analysis in different subgroups (i.e., multiple comparisons), then there are more opportunities to observe a false-positive result. When multiple comparisons are necessary, corrections (e.g., Bonferonni correction) to appropriately lower the p-value can be made in an attempt to safeguard against type I errors. A *type II error* occurs when no difference in outcomes is observed when a difference truly exists (a *false-negative* finding). That is to say, the null hypothesis was inappropriately accepted as correct. This type of error most commonly occurs when a study has insufficient power (insufficient size) to detect true differences in outcomes between groups.

Hypothesis testing can also be performed by examining confidence intervals (CI) of summary measures. Often the difference between groups are provided as an estimated ratio (in the study group divided by the control group) or as an absolute difference, with a 95% CI. The CI provides an estimate of the uncertainty around a given value. A wide CI suggests a lack of precision and a tight (small) interval indicates minimal uncertainty. When the summary measure is the absolute difference or relative risk, a CI inclusive of 0 indicates no statistically significant difference. If the summary measure is an odds ratio, a CI inclusive of 1.0 indicates no statistical difference in outcomes.

The power of a given study is the probability of rejecting the null hypothesis when it is in fact false. In more simple terms, power is a study's ability to find an association between two variables if one exists. Power is a value calculated based on a fixed and known sample size. Power is based on both study sample size and magnitude of difference observed or predicted in the dependent (response) variable in response to the independent variable. Power analysis should be done

prior to completing any statistical analysis of a study and a reasonable power for a study is widely regarded as 0.8. If the power of a study is not found to be acceptable, the reverse calculation can be made to determine a necessary sample size to determine a statistical association. Power analysis is often required for grant funding for experimental research study designs. Retrospective studies of clinical or epidemiological data with large sample sizes seldom have a preliminary power analysis.

What Are the Implications of the Study Design That Was Chosen?

Several study designs are commonly used in surgical research and depend on the study purpose (hypothesis-generating versus hypothesis-testing) and the feasibility and resources for conducting the research. A synopsis of the most common study designs in surgical literature is provided in Table 15.2—Study designs.

Randomized controlled trials (RCTs) provide the highest level of evidence supporting causality. Subjects are randomly assigned to an intervention group, where they receive an experimental intervention or to a control group, where they receive a controlled measurable alternative. If the number of randomized individuals is sufficiently large and randomization is performed properly, confounding variables will be distributed equally between groups and outcomes can be compared without concern for bias. However, conducting an RCT is challenging because of issues concerning equipoise, ethics, willingness to be randomized, costs, and generalizability.

An important analytic issue with RCTs is *intent-to-treat* (ITT). When an analysis is conducted following the ITT principle, outcome comparisons between control and treatment groups are based on the initial randomization and disregard subjects who cross over across intervention arms. If analytic approaches other than ITT are used, an equal balance of confounders across comparison groups cannot be guaranteed,

Table 15.2 Important considerations in design types

Study type	Exposure/outcome relationship	Considerations
Randomized controlled trial[T]	Randomly assigned an exposure and followed for outcome	Equipoise? Choice of control (placebo vs. standard of care) Generalizability? Blinding? Intention to treat? Superiority versus non-inferiority
Cross-sectional[T,G]	Exposure and outcome are assessed at the same point in time	Not suitable if disease has short duration or is rare
Cohort[T,G]	Identified by exposure, followed for outcome (prospective or retrospective)	One exposure, multiple outcomes Confounding Inefficient for rare outcomes or those which occur long after exposure
Case–control[T,G]	Identified by outcome, assessed for exposure (prospective or retrospective)	One outcome, multiple exposures How was control group chosen? Confounding Recall bias
Case report, series[G]		Generalizability

Adapted from Rosenthal et al. [3]
Can be considered hypothesis-testing (T) and/or hypothesis-generating (G)

and the benefits of randomization may be lost. ITT analysis is essential because it provides information about how the intervention compares at the moment the decision is being made and is particularly useful at counseling patients. When considering whether the patient should undergo a particular intervention, neither the patient nor the surgeon knows

whether the patient will be able to complete the intervention strategy or will require another approach. Instead, the ITT will communicate the intended benefit for recommending a particular intervention.

While any one study may be underpowered to answer a given research question, *meta-analysis* is a technique that pools available published data in an effort to increase the statistical power of an analysis. Meta-analysis can be applied to RCT data or observational studies. Readers should consider that guidelines have been developed to ensure the quality and validity of results obtained through RCTs, the Consolidated Standards of Reporting Trials (CONSORT) [6, 7] and meta-analysis, the QUOROM (Quality of Reporting of Meta-Analyses) [8] and MOOSE (Meta-Analysis of Observational Studies in Epidemiology) [9] meta-analysis. Regardless of the type of pooled data, in all cases, an important consideration in appraising a meta-analysis is the homogeneity of the pooled studies. Significant heterogeneity indicates more variation in study outcomes than chance alone can explain. This is particularly a concern when observational data have been aggregated because these studies tend to have less control of variability and minimal control of confounding and bias. One approach to increasing the transparency of pooled results from observational studies is to also pool the baseline characteristics of the comparison groups.

Cross-sectional studies use data collected at a single point in time and are best used for hypothesis generation. This study design is commonly used to explore relationships between variables and disease burden though the data can be stacked over time to look at temporal trends. The main limitations arise from how a population is sampled and detection or recall bias.

Cohort studies follow patients non-randomly assigned to different groups to determine whether outcomes vary across groups. While cohort data may be captured prospectively or retrospectively, the onset of observation begins with the group assignment (i.e., exposure) and continues over time to determine whether a particular event occurred. Cohort studies are useful to estimate the rates (i.e., incidence) of exposures

and outcomes, assess multiple outcomes, but are inefficient for evaluating outcomes that are rare, or occur a long time after exposure.

Case–control studies compare the frequency of exposures between patients who have and have not experienced an outcome of interest. These studies begin by enrolling subjects with and without the outcome of interest and then look back in time to search for differences in potential risk factors. Advantages of the case–control design include efficiency in evaluating the factors associated with rare outcomes or outcomes occurring a long time after exposure and the ability to evaluate multiple exposures simultaneously. Case–control designs are infrequently used in the surgical literature.

A *case report* or *series* aim to highlight an unusual or unexpected procedure or event. These studies propose a potential benefit or adverse effect of surgical therapy and may prompt more rigorous scientific evaluation. These studies are distinct from cohort investigations because there is no comparison made between competing strategies or interventions.

What Is the Variable Being Tested?

In simple terms, scientific investigation is the examination of variables. The first objective of a study is to identify the *independent* or *predictor* variable and the *dependent* or *response* variable. The dependent variable is that which changes in response to the independent variable. In experimental research, the independent variable can be manipulated to observe effects it has on the dependent variable. When it is not feasible to manipulate the independent variable for logistic, legal, or ethical reasons, nonexperimental studies attempt to show association between an independent and dependent variable through statistical inferences.

Categorical variables have discrete values and are typically described in proportions or frequencies. The simplest categorical variable is a *binary variable* that can only take on one of two values (i.e., yes/no). *Ordinal variables* are ordered

categorical variables (i.e., ASA class). *Nominal variables* are unordered categorical variables (i.e., ethnicity). A *continuous variable* is one that can take on any number of values within a specified range of possibilities. Age is an example of a continuous variable. Descriptive statistics are used to describe the central tendency of continuous variables. The arithmetic mean provides a good estimate of central tendency for normally distributed (Gaussian or bell-shaped) data. If the data are skewed (not normally distributed), the mean will be a biased estimator of the central tendency. In these cases, the median or geometric mean provides a better estimate.

Time-to-event variables consist of two variables, a continuous variable that measures the time interval from an established start point (e.g., date of diagnosis or therapy) to a binary failure event (e.g., death or disease recurrence) or the end of the observation period. Time-to-event variables are typically reported as a probability of an event occurring at a certain point in time (i.e., survival at 5 years). Typically, time-to-event methods (the most commonly used is Kaplan-Meier) consider that number of patients at risk for an event decreases over time. Because of this, some methods may overestimate risk in the setting of competing risks (the disease evolves and prompts re-intervention; over time, a contraindication to re-intervention may develop, or death may occur, in which case a patient is no longer at risk). However, methods exist for handling time to event variables in the setting of competing risks [10].

Was the Correct Analysis Performed?

Central Tendency

The point at which observations tend to cluster is a frequent point of interest in scientific investigation. The mean, median, and mode of a group of observations each provide an assessment of this tendency and have their own practical uses. The *mean* is the summation of all observations for a given group

of data divided by the number of observations in that data. The mean is highly sensitive to outlying observations within a dataset and can be an invalid assessment of central tendency if the data are skewed to one direction. The *median* is not as influenced by outlying observations and is defined as the observation at the 50th percentile for a group of observations. The third most commonly used and reported measure of central tendency is the *mode*, which is the set of values in a group of observations that occurs most frequently.

When considering a measure of central tendency, it is also important to consider the dispersion of the observations around the measure. The *range* of a group of measurements is the difference between the largest and smallest observation in a dataset and can give a crude assessment of the dispersion of observations. This value is again heavily influenced by outlying observations. The *variance* of a set of observations is the sum of squared distance from all observations to the mean in a given group of observations divided by 1-number of observations, and *standard deviation* is the square root of the variance. Standard deviation is the most widely reported assessment of dispersion due to its properties. For a relatively symmetric group of data, 67% of the observations will be within +/− one standard deviation from the mean and 95% of the observations will be within +/− two standard deviations from the mean. The 95% confidence interval of a mean is thus bounded by the values two standard deviations below and above the mean, respectively. Furthermore, a mean from one group of observations can be said to be significantly different from a mean of another set of observations if the 95% confidence intervals do not overlap.

Rates

Some demographic data describing a population can be reported through measures of central tendency such as age and BMI. Not all data can be reported using these measures. Many demographic variables such as gender, race, comorbidities, and behavioral attributes such as smoking must be

reported as rates. Other vital statistics such as births, deaths, and disease prevalence and incidence are also reported as rates.

Probability

An extension of rates is probability. The relative risk (RR) is a comparison of probabilities. RR is the probability of an event such as death, disease, or complication in subjects with a given exposure compared to the probability of death, disease, or complication in subjects without this exposure. A relative risk of 1.0 would indicate that the risk for death among the diseased group would be identical to those without disease. The odds ratio is a comparison of the odds as opposed to the probability of an event. While probability is proportion of outcome of interest to all observations, odds are the outcome of interest in proportion to the alternative outcome. The odds of an event are not a risk or probability or risk per se. As such, it is a more appropriate statistic to compute in retrospective studies such as case–control or cross-sectional studies, where risk cannot be determined. In prospective cohort studies and RCTs, relative risk is an acceptable statistic.

It is important to note that the odds ratio will overestimate the probability if the outcome occurs frequently (>10%) in the population [11]. When the outcome is rare, the odds generally provide a good approximation of the probability. It is particularly relevant when conducting multivariable analysis that a minimum number of events are included to achieve a reliable estimate. As rules of thumb, a minimum of 10 events (and equivalent number of nonevents) per variable are required for logistic regression (binary outcome) [12] and 10–15 observations per variable for linear regression (continuous outcome) [13]. This should also be considered when multiple variables are being controlled for in a multivariable model.

Diagnostic Testing

Probability forms the basis for the value of a diagnostic test. The probability of disease given a test result is paramount to accurately diagnosing or ruling out disease in a patient. Multiple measures of probability are used to assess the accuracy of a given diagnostic test. The *sensitivity* of a test is the probability of a positive test given a patient has the disease being tested. Another assessment of accuracy is *specificity* which is the probability of a negative test given a patient does not have the disease. These are important measures of accuracy for establishing the usefulness of a screening tool. Many times an individual will wonder what the probability of disease is given they test positive or the probability that they don't have disease given a negative test result. These measures are the *positive* and *negative predictive values,* respectively. Positive and negative predictive values are heavily influenced by disease prevalence while sensitivity and specificity are not. As a result of this variation in diagnostic power, PPV and NPV are less favored to the positive and negative likelihood ratios as both of these tests can be calculated using sensitivity and specificity. The *positive likelihood ratio* can be calculated as (sensitivity)/(1-specificity) and the *negative likelihood ratio* can be calculated as (1-sensitivity)/(specificity). Positive and negative likelihood ratios greater than 10 and less than 0.1, respectively, offer significant shifts in likelihood of disease. Positive and negative likelihood ratios less than 2 and greater than 0.5, respectively, do not suggest significant impact on likelihood of diagnosis.

Statistical Testing

The *Student's t-test* is one of the most common tests for analyzing sample means. The *t*-test is based on null and alternative hypotheses. Notwithstanding the type of *t*-test being run, the null hypothesis is always that no difference exists. A one sample *t*-test is used to test if a sample mean is different from

a known population mean. A two-sample *t*-test can be either paired or independent. A paired, two-sample *t*-test is used to test the difference between matched samples, such as the mean systolic blood pressure in patients before starting a drug and those same patients after 6 months on therapy. An independent, two sample *t*-tests are used to test the difference in sample means between two unrelated samples, such as the difference in mean systolic blood pressure between patients with diabetes and patients without. A one-sided *t*-test is used if it is known that the test sample will either be less than or greater than a reference point. A two-sided *t*-test is used if this is not necessarily known.

The Student's *t*-test is limited in its utility to two samples. If a study design calls for multiple samples such as repeated measurements on a sample group or sampling from more than two groups, repeatedly running the Student's *t*-test increases the chance of a type I error, or finding an association by chance. In this setting, *Analysis of Variance (ANOVA)* is a more appropriate test. The ANOVA as its name would suggest creates a test statistic from of the dispersion or variance of a variable. The term variance in this instance is referring to general dispersion of the sample data, rather than the variance value described above. For the ANOVA, the dispersion or variance is calculated using the sum of squares method, which is again beyond the scope of this chapter. This variance (dispersion) is partitioned into, or calculated for, all subjects within-groups and between-groups. In this way, a single test statistic termed the F-statistic can be used to effectively determine if the means of three or more groups are different from each other. The F-statistic also correlates to a probability that is determined from the *F* distribution and this probability is again, the p-value. In a one-way ANOVA, the variation in the response variable is attributed to a single factor, i.e., the difference of the different sample groups. For example, suppose a research team is interested in recurrence of Crohn's ileitis after surgery and the effects of multiple types of drugs on remission, such as budesonide, methotrexate, and infliximab. The different drug regimens a subject is

taking is the source of variation. A two-way ANOVA can be done if there are two factors such as anti-inflammatory medications and diet routines that could potentially affect recurrence. The two-way ANOVA will give a statistic for both medications and diet and the interaction of these two factors. The important thing to remember with ANOVA is that it will not indicate which groups are significantly different from each other, only that there is significant variation by group. Two group means may be the same, statistically speaking, while only the third is significantly different. In this instance, a *t*-test may be used to determine which group mean is different. However, it is important to again state that running multiple *t*-tests will raise the likelihood of a type I error. This can be overcome by lowering the threshold for significance below 0.05–0.01 or even lower. Additionally, multivariable methods can be employed to determine the individual group effects on the response variable.

The above tests are designed to assess the differences in group means. Often times, a study doesn't collect data that can be reported as a mean. As described above, the response variable is a yes or a no, disease or no disease, complication or no complication. Also, as described above, this can be described in terms of probability, risk, and odds. An alternative means of assessing categorical data of this nature is the *chi-square test*. The chi-square test makes use of contingency tables which, in their simplest form, are 2×2 tables with dichotomous dependent and independent variables. A 2×2 table could be constructed for patients who undergo cholecystectomy. Columns are patients who underwent either laparoscopic or open procedures and rows are bile leak versus no bile leak. The odds ratio or relative risk could be calculated in this instance. The chi-square test can also tell if the rate of bile leak is different between the two procedure types. The benefit of the chi-square test is that it can be extended to categorical data with more than two responses. For example, rates of bile leak could be compared to ASA class which has six responses. The chi-square test can be used to determine if rates of bile leak are increased based for patients with higher

ASA classifications. The chi-square test again utilizes a test statistic calculated based on the observed and expected counts for each cell in the contingency table and a p-value is obtained from the chi-square distribution. A test called McNemar's test can be employed when data are paired.

Regression is a tool that is useful predicting response from some predictor variable. There are multiple regression methods that can be applied to different variable types. A *simple linear regression* is the most basic example and can be used when the response variable and predictor variable are continuous. For example, blood pressure changes with changes in subject age can be evaluated in this way. Regression makes use of the variation in response variables as they relate to the dependent variable. Another way to say this is, it uses the average of the response variable when the predictor value is fixed (i.e., the BP ranges for patients at age 35). If this average changes significantly when the predictor variable is changed, the predictor variable is a significant predictor of the response variable. Mathematics is unimportant as they are typically done by statistical analysis programs. With that said, a *p*-value is obtained that gives the probability that the variation in response due to the predictor is seen by chance. *Multivariable linear regression* can be utilized when there are multiple predictor variables being assessed. This is not to be confused with multivariate regression which is used for multiple response or dependent variables. Regression can also be performed on dichotomous or categorical outcome variables. In this instance, the logistic regression model is used. The logistic regression model will also provide a p-value for the association between the predictor and response but it will also provide an odds ratio for the given predictor variable.

All of the above tests rely on multiple assumptions for their validity. One of the most significant assumptions is that the parameter, or numerical characteristic of the population from which the study sample is drawn, fits a specified distribution. Typically, the assumption is a normal distribution. The central limit theorem makes this assumption true for most variables. However, this is not always the case. Nonparametric

tests do not require this assumption, but have similar procedures to the above tests making use of other statistics such as the median. The *Wilcoxon Rank Sum* is the nonparametric equivalent of the Student's *t*-test and makes use of group medians rather than means and the *Kruskal–Wallis test* is the nonparametric equivalent of the ANOVA. Additionally, data transformations can be done, such as log(Variable) or ln(Variable) that can give the sample data a normal distribution. The decision to transform data and the method of transformation should be decided on during the design phase of a study and not part of post hoc analysis.

Table 15.3 summarizes the types of tests according to the categories of variables used.

How to Interpret the Study Findings?

One of the most important issues to consider in the evaluation and conduct of outcomes research, especially when observational data is being used, is *confounding*. A confounder is a measured or unmeasured variable associated with the exposure of interest and associated with the outcome. This dual relationship can influence the degree and direction of, or even completely mitigate, an observed association between exposure and outcome [14]. RCTs can address confounding through randomization. On the other hand, investigators who perform observational studies must address confounding both with analytical approaches (i.e., multivariable regression and propensity scores), and acknowledge potential residual confounding in their discussion of the study's limitations, noting variables that were not measured, their relationship with the exposure and the outcome, and their implication on the potential direction and magnitude of confounding bias.

It is important to consider the many forms of bias when evaluating research. For instance, there any many forms of selection bias which can favor administration of a particular intervention to those thought to need it the most. *Propensity score* analysis is an alternative method of risk adjustment to

Table 15.3 Summary of statistical testing by variable type

Outcome→	**Continuous**		**Categorical**	
Exposure↓	**Dichotomous**		**Polychotomous**	
			Ordinal	**Nominal**
Continuous Univariate	Pearson Correlation Linear regression *Spearman rank correlation*	Logistic regression Discriminant analysis *Wilcoxon rank sum*	Ordinal regression *Spearman rank correlation*	Discriminant analysis *Kruskal–Wallis*
Multivariate	Multiple linear regression	Logistic regression Discriminant analysis	Ordinal regression	Discriminant analysis Nominal regression
CATEGORICAL ***Dichotomous*** Univariate	*t*-test ANOVA Linear regression *Wilcoxon rank sum*	Chi-square Logistic regression	*Wilcoxon rank sum* Chi-square test for trend Ordinal regression	Chi-square
Multivariate	*N*-way ANOVA Analysis of covariance Multiple linear regression	Mantel–Haenszel Logistic regression	Ordinal regression	Discriminant analysis Nominal regression

Matched	Matched pairs *t*-test	NcNemar's test (univariate) Conditional logistic regression (multivariate)	Does not exist	Does not exist
Polychotomous ordinal or nominal Univariate	ANOVA Linear regression *Spearman rank correlation (ordinal)* *Kruskal–Wallis (nominal)*	Chi-square Logistic regression *Spearman rank correlation (ordinal)*	Chi-square test for trend Ordinal regression *Spearman rank correlation*	Chi-square Discriminant analysis Nominal regression
Multivariate	*N*-way ANOVA Multiple regression	Logistic regression	Ordinal regression	Chi-square Discriminant analysis Nominal regression
Matched	Repeated measures ANOVA Multiple linear regression	Conditional logistic regression	Does not exist	Does not exist

Italic text = nonparametric tests

reduce the bias in estimating treatment effects when analyzing nonrandomized, observational data. Because patients receiving one treatment tend to be different than patients receiving another (e.g., minimally invasive [MIS] as compared with open surgery), a propensity score is calculated using logistic regression to determine a subject's probability of having the exposure of interest (propensity to undergo MIS). The outcomes of interest for patients who do and do not undergo MIS (but have a similar propensity to undergo MIS) can then be compared through matching, stratified analyses, or regression (adjusting only for propensity) [15].

While propensity score analysis cannot adjust away all confounders, known or unknown, as these are intrinsic to observational data. However, there are three circumstances in which the use of propensity scores may be appropriate: (1) there are many confounders relative to the number of events (i.e., less than ten events per covariate) resulting in an unpowered regression analysis; (2) there is no interest in the association between the adjustment factors and outcome; and (3) the relationship between the exposure and propensity for treatment can be estimated more accurately than the relationship between the exposure and outcome [16].

Generalizability refers to the application of research findings to routine, clinical practice. While RCTs provide the highest level of evidence about the efficacy of competing interventions, they are conducted in a highly controlled environment, limiting other providers' ability to reproduce the delivery of care and outcomes in a non-research setting. In addition, generalizability issues apply to observational studies as well. Critical readers should consider why care patterns and outcomes described in research studies might not be reproducible in other clinical settings and patient populations.

Conclusions

Although not all jobs in surgery require academic productivity in the form of original clinical research, a basic familiarity with data sources, study design, and statistical testing can

greatly enhance a young surgeon's ability to intelligently contribute to discussions involving the scientific literature. Early in our careers, this proficiency can manifest in many beneficial ways that stretch well beyond the surgical journal club setting: engaging our peers in multidisciplinary conferences, incorporating evidence-based principles into our own practices, volunteering to adjudicate scientific abstracts for society meetings, and providing peer-review for editorial boards, to name a few. We encourage you to approach any review with these five questions in mind: (1) What is the study's purpose? (2) Is the right data being used? (3) What is the measured outcome? (4) What are the implications of the study design that was chosen? (5) Was the correct analysis performed (and adequately powered to support the conclusion)?

References

1. Richardson WS, Wilson MC, Nishikawa J, Hayward RS. The well-built clinical question: a key to evidence-based decisions. ACP J Club. 1995;123(3):A12–3.
2. Schardt C, Adams MB, Owens T, Keitz S, Fontelo P. Utilization of the PICO framework to improve searching PubMed for clinical questions. BMC Med Inform Decis Mak. 2007;7:16.
3. Rosenthal R, Schafer J, Briel M, Bucher HC, Oertli D, Dell-Kuster S. How to write a surgical clinical research protocol: literature review and practical guide. Am J Surg. 2014;207(2):299–312.
4. Fleming TR, Powers JH. Biomarkers and surrogate endpoints in clinical trials. Stat Med. 2012;31(25):2973–84.
5. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials. 1989;10(4):407–15.
6. Turner L, Shamseer L, Altman DG, Weeks L, Peters J, Kober T, et al. Consolidated standards of reporting trials (CONSORT) and the completeness of reporting of randomised controlled trials (RCTs) published in medical journals. Cochrane Database Syst Rev. 2012;11:MR000030.
7. Nagendran M, Harding D, Teo W, Camm C, Maruthappu M, McCulloch P, et al. Poor adherence of randomised trials in surgery

to CONSORT guidelines for non-pharmacological treatments (NPT): a cross-sectional study. BMJ Open. 2013;3(12):e003898, 2013–003898.
8. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. quality of reporting of meta-analyses. Lancet. 1999;354(9193):1896–900.
9. Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. meta-analysis of observational studies in epidemiology (MOOSE) group. JAMA. 2000; 283(15):2008–12.
10. Resche-Rigon M, Azoulay E, Chevret S. Evaluating mortality in intensive care units: contribution of competing risks analyses. Crit Care. 2006;10(1):R5.
11. Chen W, Shi J, Qian L, Azen SP. Comparison of robustness to outliers between robust poisson models and log-binomial models when estimating relative risks for common binary outcomes: a simulation study. BMC Med Res Methodol. 2014;14:–82. doi:10.1186/1471-2288-14-82.
12. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996;49(12):1373–9.
13. Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosom Med. 2004;66(3):411–21.
14. Mehio-Sibai A, Feinleib M, Sibai TA, Armenian HK. A positive or a negative confounding variable? A simple teaching aid for clinicians and students. Ann Epidemiol. 2005;15(6):421–3.
15. Haukoos JS, Lewis RJ. The propensity score. JAMA. 2015;314(15):1637–8.
16. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivar Behav Res. 2011;46(3):399–424.
17. Simianu VV, Farjah F, Flum D. Evidence-based surgery: critically assessing surgical literature (Chapter 8). In: Townsend CM, Beauchamp RD, Evers BM, Mattox KL, editors. Sabiston textbook of surgery. Philadelphia: Elsevier; 2016.