

CURRENT CONCEPTS REVIEW

CLINICAL EPIDEMIOLOGY AND BIostatISTICS: A PRIMER FOR ORTHOPAEDIC SURGEONS

BY MININDER S. KOCHER, MD, MPH, AND DAVID ZURAKOWSKI, PHD

Investigation performed at Harvard Medical School, Harvard School of Public Health, Boston, Massachusetts

Epidemiology is the study of the distribution and determinants of disease frequency¹. In the fifth century BC, Hippocrates suggested that the development of human disease might be related to the external and internal environment of an individual¹. In the 1600s and 1800s in England, John Graunt and William Farr quantified vital statistics on the basis of birth and death records¹. In the 1850s, John Snow associated cholera with water contamination in London by observing higher cholera rates in homes supplied by certain water sources¹. Epidemiological methods gradually evolved with use of the case-control study to demonstrate an association between smoking and lung cancer, use of the prospective cohort study to determine risk factors for cardiovascular disease in the Framingham Heart Study, and use of the randomized clinical trial for the poliomyelitis vaccine¹. The evidence-based medicine and patient-derived outcomes assessment movements burst onto the scene of clinical medicine in the 1980s and 1990s as a result of contemporaneous medical, societal, and economic influences. Pioneers such as Sackett and Feinstein emphasized levels of evidence and patient-centered outcomes assessment²⁻¹⁰. Work by Wennberg and colleagues revealed large small-area variations in clinical practice, with some patients being thirty times more likely to undergo an operative procedure than other patients with identical symptoms merely because of their geographic location¹¹⁻¹⁶. Additional critical research suggested that up to 40% of some surgical procedures might be inappropriate and up to 85% of common medical treatments were not rigorously validated¹⁷⁻¹⁹. Meanwhile, the costs of health care were rapidly rising to over two billion dollars per day, increasing from 5.2% of the gross domestic product in 1960 to 16.2% in 1997²⁰. Health maintenance organizations and managed care emerged. In addition, increasing federal, state, and consumer oversight was brought to bear on the practice of clinical medicine.

These forces have led to an increased focus on the effectiveness of clinical care. Clinical epidemiology provides the methodology with which to assess this effectiveness. This article presents an overview of the concepts of study design, hypothesis testing, measures of treatment effect, diagnostic

performance, evidence-based medicine, outcomes assessment, data, and statistical analysis. Examples from the orthopaedic literature and a glossary of terminology (terms italicized throughout the text) are provided.

Study Design

In *observational studies* researchers observe patient groups without allocation of the intervention, whereas in *experimental studies* researchers allocate the treatment. Experimental studies involving humans are called trials. Research studies may be *retrospective*, meaning that the direction of inquiry is backward from the cases and that the events of interest transpired before the onset of the study. Alternatively, studies may be *prospective*, meaning that the direction of inquiry is forward from the cohort inception and that the events of interest transpire after the onset of the study (Fig. 1). *Cross-sectional* studies are used to survey one point in time. *Longitudinal* studies follow the same patients over multiple points in time.

All research studies are susceptible to invalid conclusions due to bias, confounding, and chance. *Bias* is the non-random systematic error in the design or conduct of a study. Bias usually is not intentional; however, it is pervasive and insidious. Forms of bias can corrupt a study at any phase, including patient selection (selection and membership bias), study performance (performance and information bias), patient follow-up (nonresponder and transfer bias), and outcome determination (detection, recall, acceptability, and interviewer bias). Frequent biases in the orthopaedic literature include selection bias, when dissimilar groups are compared; nonresponder bias, when the follow-up rate is low; and interviewer bias, when the investigator determines the outcome. A *confounder* is a variable that has independent associations with both the *independent* (predictor) and *dependent* (outcome) variables, thus potentially distorting their relationship. For example, an association between knee laxity and anterior cruciate ligament injury may be confounded by female sex since women may have greater knee laxity and a higher risk of anterior cruciate ligament injury. Frequent confounders in clinical research include gender, age, socioeconomic status, and co-

morbidities. As discussed below in the section on hypothesis testing, chance may lead to invalid conclusions based on the probability of *type-I* and *type-II* errors, which are related to *p* values and *power*.

The adverse effects of bias, confounding, and chance can be minimized by study design and statistical analysis. Prospective studies minimize bias associated with patient selection, quality of information, attempts to recall preoperative status, and nonresponders. *Randomization* minimizes selection bias and equally distributes confounders. *Blinding* can further decrease bias, and *matching* can decrease confounding. Confounders can sometimes be controlled post hoc with the use of stratified analysis or multivariate methods. The effects of chance can be minimized by an adequate sample size based on power calculations and use of appropriate levels of significance in hypothesis testing. The ability of study design to optimize validity while minimizing bias, confounding, and chance is recognized by the adoption of hierarchical levels of evidence on the basis of study design (see Table [Levels of Evidence for Primary Research Question] in Instructions to Authors of this issue of *The Journal*). Furthermore, the standard to prove cause-effect is set higher than the standard to suggest an association. Inference of causation requires supporting data from non-observational studies such as a randomized clinical trial, a biologically plausible explanation, a relatively large effect size, reproducibility of findings, a temporal relationship between cause and effect, and a biological gradient demonstrated by a dose-response relationship.

Observational study designs include case series, case-control studies, cross-sectional surveys, and cohort studies. A *case series* is a retrospective, descriptive account of a group of patients with interesting characteristics or a series of patients who have undergone an intervention. A case series that includes one patient is a case report. Case series are easy to construct and can provide a forum for the presentation of interesting or unusual observations. However, case series are often anecdotal, are subject to many possible biases, lack a hypothesis, and are difficult to compare with other series. Thus, case series are usually viewed as a means of generating hypotheses for addi-

tional studies but not as conclusive. A *case-control study* is a study in which the investigator identifies patients with an outcome of interest (cases) and patients without the outcome (controls) and then compares the two groups in terms of possible risk factors. The effects in a case-control study are frequently reported with use of the *odds ratio*. Case-control studies are efficient (particularly for the evaluation of unusual conditions or outcomes) and are relatively easy to perform. However, an appropriate control group may be difficult to identify, and preexisting high-quality medical records are essential. Moreover, case-control studies are susceptible to multiple biases, particularly selection and detection biases based on the identification of cases and controls. Cross-sectional surveys are often used to determine the prevalence of disease or to identify coexisting associations in patients with a particular condition at one particular point in time. The *prevalence* of a condition is the number of individuals with the condition divided by the total number of individuals at one point in time. *Incidence*, in contradistinction, refers to the number of individuals with the condition divided by the total number of individuals over a defined time period. Thus, prevalence data are usually obtained from a cross-sectional survey creating a proportion, whereas incidence data are usually obtained from a prospective cohort study and a time value is contained in the denominator. Surveys are also frequently performed to determine preferences and treatment patterns. Because cross-sectional studies represent a snapshot in time, they may be misleading if the research question involves the disease process over time. Surveys also present unique challenges in terms of adequate response rate, representative samples, and acceptability bias. A traditional *cohort study* is one in which a population of interest is identified and is followed prospectively in order to determine outcomes and associations with risk factors. Retrospective, or historical, cohort studies can also be performed; in those studies, cohort members are identified on the basis of records, and the follow-up period is entirely or partly in the past. Cohort studies are optimal for studying the incidence, course, and risk factors of a disease because they are longitudinal, meaning that a group of subjects is followed over time.

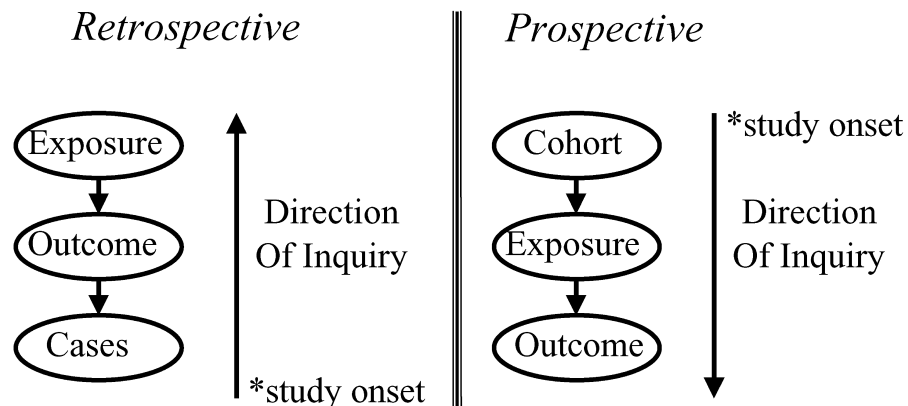


Fig. 1

Comparison of prospective and retrospective study designs on the basis of the direction of inquiry and the onset of the study.

TABLE I Hypothesis Testing*

Experiment	Truth	
	No Association	Association
No association	Correct	Type-II (β) error
Association	Type-I (α) error	Correct

*P value = probability of type-I (α) error. Power = 1 – probability of type-II (β) error.

The effects in a cohort study are frequently reported in terms of *relative risk* (RR). Because traditional cohort studies are prospective, they can optimize follow-up and data quality and can minimize bias associated with selection, information, and measurement. In addition, they have the correct time-sequence to provide strong evidence regarding associations. However, these studies are costly, are logistically demanding, often require a long time-period for completion, and are inefficient for the assessment of unusual outcomes or diseases.

Experimental study designs may involve the use of concurrent controls, sequential controls (*crossover trials*), or historical controls. The *randomized clinical trial* (RCT) with concurrent controls is the so-called gold standard of clinical evidence as it provides the most valid conclusions (internal validity) by minimizing the effects of bias and confounding. Rigorous randomization with enough patients is the best means of avoiding confounding. The performance of a randomized control trial involves the construction of a protocol document that explicitly establishes eligibility criteria, sample size, informed consent, randomization, rules for stopping the trial, blinding, measurement, monitoring of compliance, assessment of safety, and data analysis. Because allocation is random, selection bias is minimized and confounders (known and unknown) theoretically are equally distributed between groups. Blinding minimizes performance, detection, interviewer, and acceptability bias. Blinding may be practiced at four levels: participants, investigators applying the intervention, outcome assessors, and analysts. *Intention-to-treat analysis* minimizes nonresponder and transfer bias, while sample-size determination ensures adequate power. The intention-to-treat principle states that all patients should be analyzed within the treatment group to which they were randomized in order to preserve the goals of randomization. Although the randomized clinical trial is the epitome of clinical research designs, the disadvantages of such trials include their expense, logistics, and time to completion. Accrual of patients and acceptance by clinicians may be difficult. With rapidly evolving technology, a new technique may quickly become well accepted, making an existing randomized clinical trial obsolete or a potential randomized clinical trial difficult to accept. Ethically, randomized clinical trials require clinical equipoise (equality of treatment options in the clinician's judgment) for enrollment, interim stopping rules to avoid harm and to evaluate adverse events, and truly informed consent. Finally, while

randomized clinical trials have excellent internal validity, some have questioned their generalizability (external validity) because the practice pattern and the population of patients enrolled in a randomized clinical trial may be overly constrained and nonrepresentative.

Ethical considerations are intrinsic to the design and conduct of clinical research studies. Informed consent is of paramount importance, and it is the focus of much of the activity of institutional review boards. Investigators should be familiar with the Nuremberg Code and the Declaration of Helsinki as they pertain to ethical issues of risks and benefits, protection of privacy, and respect for autonomy^{21,22}.

Hypothesis Testing

The purpose of hypothesis testing is to permit generalizations from a *sample* to the population from which it came. Hypothesis testing confirms or refutes the assertion that the observed findings did not occur by chance alone but rather occurred because of a true association between variables. By default, the *null hypothesis* of a study asserts that there is no significant association between variables whereas the alternative hypothesis asserts that there is a significant association. If the findings of a study are not significant we cannot reject the null hypothesis, whereas if the findings are significant we can reject the null hypothesis and accept the alternative hypothesis.

Thus, all research studies that are based on a sample make an inference about the truth in the overall population. By constructing a 2×2 table of the possible outcomes of a study (Table I), we can see that the inference of a study is correct if a significant association is not found when there is no true association or if a significant association is found when there is a true association. However, a study can have two types of errors. A *type-I* or *alpha* (α) error occurs when a significant association is found when there is no true association (resulting in a false-positive study that rejects a true null hypothesis). A *type-II* or *beta* (β) error wrongly concludes that there is no significant association (resulting in a false-negative study that rejects a true alternative hypothesis).

The alpha level refers to the probability of a type-I (α) error. By convention, the alpha level of significance is set at 0.05, which means that we accept the finding of a significant association if there is less than a one in twenty chance that the observed association was due to chance alone. Thus, the p value, which is calculated with a statistical test, is a measure of the strength of the evidence provided by the data in favor of the null hypothesis. If the p value is less than the alpha level, then the evidence against the null hypothesis is strong enough for us to reject it and conclude that the result is significant.

P values frequently are used in clinical research and are given great importance by journals and readers; however, there is a strong movement in biostatistics to deemphasize p values because a significance level of $p < 0.05$ is arbitrary, a strict cut-off point can be misleading (there is little difference between $p = 0.049$ and $p = 0.051$, but only the former is considered "significant"), the p value gives no information about the strength of the association, and the p value may be statistically signifi-

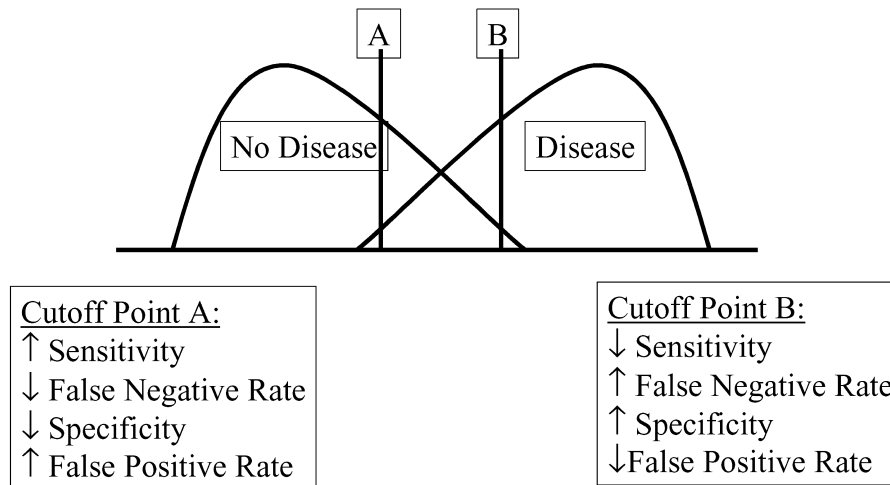


Fig. 2

Selection of positivity criterion. Because there is typically overlap between the diseased population and the nondiseased population over a range of diagnostic values (x axis), there is an intrinsic trade-off between sensitivity and specificity. When the positive test results are identified as those to the right of cutoff point A, there is high sensitivity because most patients with the disease are correctly identified as having a positive result. However, there is lower specificity because some of the patients without the disease are incorrectly identified as having a positive result (false positives). When the positive test results are identified as those to the right of cutoff point B, there is lower sensitivity because some patients with the disease are incorrectly identified as having a negative result (false negatives). However, there is high specificity because most patients without the disease are correctly identified as having a negative result.

cant without the results being clinically important. Alternatives to the traditional reliance on p values include the use of variable alpha levels of significance based on the consequences of the type-I error and the reporting of p values without using the term “significant.” Use of 95% *confidence intervals* in lieu of p values has gained acceptance as these intervals convey information regarding the significance of findings (95% confidence intervals do not overlap if they are significantly different), the magnitude of differences, and the precision of measurement (indicated by the range of the 95% confidence interval). Whereas the p value is often interpreted as being either significant or not, the 95% confidence interval provides a range of values that allows the reader to interpret the implications of the results. In addition, while p values have no units, confidence intervals are presented in the units of the variable of interest, which helps the reader to interpret the results. For example, the authors of a study of the duration of the hospital stay for children with septic arthritis of the hip managed according to a clinical practice guideline may state that “there was a significantly shorter hospital stay for patients treated according to the guideline” with the addition of either “ $p = 0.003$ ” if p values are used or “95% confidence intervals, 3.8 to 5.8 days for patients treated according to the guideline and 7.3 to 9.3 days for patients not treated according to the guideline” if 95% confidence intervals are used²³. The p-value approach conveys statistical significance only, whereas the confidence-interval approach conveys statistical significance (the confidence intervals do not

overlap), clinical significance (the magnitude of the values), and precision (the range of the confidence intervals).

Power is the probability of finding a significant association if one truly exists and is defined as $1 - \text{the probability of a type-II } (\beta) \text{ error}$. By convention, acceptable power is set at $\geq 80\%$, which means that there is $\leq 20\%$ chance that the study will demonstrate no significant association when there is a true association. In practice, when a study demonstrates a significant association, the potential error of concern is the type-I (α) error as expressed by the p value. However, when a study demonstrates no significant association, the potential error of concern is the type-II (β) error as expressed by power—that is, in a study that demonstrates no significant effect, there may truly be no significant effect or there may actually be a significant effect but the study was underpowered because the sample size was too small or the measurements were too imprecise. Thus, when a study demonstrates no significant effect, the power of the study should be reported.

The calculations for power analyses differ depending on the statistical methods that are utilized for the analysis; however, four elements are involved in a power analysis: α , β , effect size, and sample size (n). Effect size is the difference that you want to be able to detect with the given α and β . It is based on a clinical sense about how large a difference would be clinically meaningful. Effect sizes are often defined in dimensionless terms, on the basis of a difference in mean values divided by the pooled standard deviation for a comparison of two

groups. Small sample sizes, small effect sizes, and large variances all decrease the power of a study. An understanding of power issues is important in clinical research, to minimize the use of resources when planning a study and to ensure the validity of a study. Sample-size calculations are performed when a study is being planned. Typically, power is set at 80%, alpha is set at 0.05, the effect size and variance are estimated from pilot data or the literature, and the equation is solved for the necessary sample size. Calculation of power after the study has been completed—that is, post-hoc power analysis—is controversial and is discouraged.

Diagnostic Performance

A diagnostic test can result in four possible scenarios: (1) true positive if the test is positive and the disease is present, (2) false positive if the test is positive and the disease is absent, (3) true negative if the test is negative and the disease is absent, and (4) false negative if the test is negative and the disease is present (Table II). The *sensitivity* of a test is the percentage (or proportion) of patients with the disease who are classified as having a positive result of the test (the true-positive rate). A test with 97% sensitivity implies that, of 100 patients with the disease, ninety-seven will have a positive test. Sensitive tests have a low false-negative rate. A negative result of a highly sensitive test rules disease out (SNout). The *specificity* of a test is the percentage (or proportion) of patients without the disease who are classified as having a negative result of the test (the true-negative rate). A test with 91% specificity implies that, of 100 patients without the disease, ninety-one will have a negative test. Specific tests have a low false-positive rate. A positive result of a highly specific test rules disease in (SPin). Sensitivity and specificity can be combined into a single parameter, the *likelihood ratio (LR)*, which is the probability of a true positive divided by the probability of a false positive. Sensitivity and specificity can be established in studies in which the results of a diagnostic test are compared with those of the “gold standard” of diagnosis in the same patients—for example, by comparing the results of magnetic resonance imaging with arthroscopic findings²⁴.

Sensitivity and specificity are technical parameters of diagnostic testing performance and have important implications for screening and clinical practice guidelines^{25,26}; however, they are less relevant in the typical clinical setting because the clinician does not know whether the patient has the disease. The

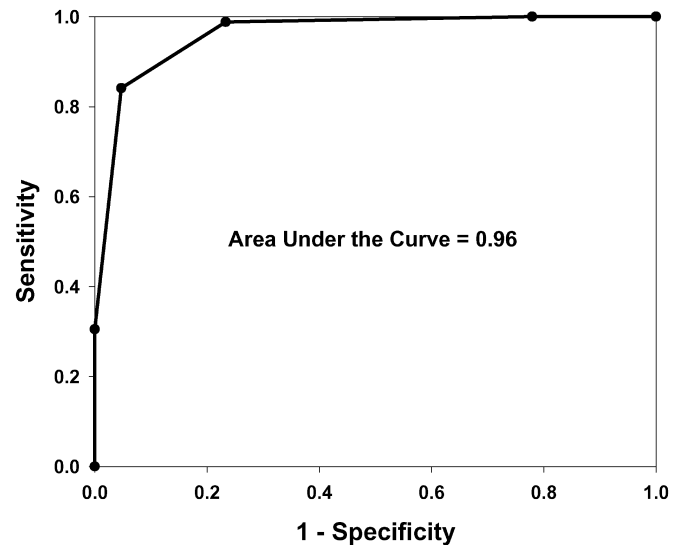


Fig. 3

Receiver-operating characteristic (ROC) curve for a clinical prediction rule for differentiating septic arthritis from transient synovitis of the hip in children²⁸. The false-positive rate (1 – specificity) is plotted on the x axis, and sensitivity is plotted on the y axis. The area under the curve represents the overall diagnostic performance of a prediction rule or a diagnostic test. For a perfect test, the area under the curve is 1.0. For random guessing, the area under the curve is 0.5.

clinically relevant issues are the probability of the patient having the disease when the result is positive (*positive predictive value [PPV]*) and the probability of the patient not having the disease when the result is negative (*negative predictive value [NPV]*). The positive and negative predictive values are probabilities that require an estimate of the prevalence of the disease in the population, and they can be calculated with use of equations that utilize Bayes' theorem²⁷.

There is an inherent trade-off between sensitivity and specificity. Because there is typically some overlap between the diseased and nondiseased groups with respect to a test distribution, the investigator can select a positivity criterion with a low false-negative rate (to optimize sensitivity) or a low false-positive rate (to optimize specificity) (Fig. 2). In practice, positivity criteria are selected on the basis of the consequences of a false-positive or a false-negative diagnosis. If the consequences of a false-negative diagnosis outweigh the consequences of a

TABLE II Diagnostic Test Performance*

	Disease Positive	Disease Negative
Test positive	a (true positive)	b (false positive)
Test negative	c (false negative)	d (true negative)

*Sensitivity = $a/(a + c)$, specificity = $d/(b + d)$, accuracy = $(a + c)/(a + b + c + d)$, false-negative rate = $1 - \text{sensitivity}$, false-positive rate = $1 - \text{specificity}$, likelihood ratio (+) = $\text{sensitivity}/\text{false-positive rate}$, likelihood ratio (-) = $\text{false-negative rate}/\text{specificity}$, positive predictive value = $[(\text{prevalence})(\text{sensitivity})]/[(\text{prevalence})(\text{sensitivity}) + (1 - \text{prevalence})(1 - \text{specificity})]$, and negative predictive value = $[(1 - \text{prevalence})(\text{specificity})]/[(1 - \text{prevalence})(\text{specificity}) + (\text{prevalence})(1 - \text{sensitivity})]$.

TABLE III Treatment Effects*

	Adverse Events	No Adverse Events
Experimental group	a	b
Control group	c	d

*Control event rate (CER) = $c/(c + d)$, experimental event rate (EER) = $a/(a + b)$, control event odds (CEO) = c/d , experimental event odds (EEO) = a/b , relative risk (RR) = EER/CER , odds ratio (OR) = EEO/CEO , relative risk reduction (RRR) = $(EER - CER)/CER$, absolute risk reduction (ARR) = $EER - CER$, and number needed to treat (NNT) = $1/ARR$.

false-positive diagnosis of a condition (such as septic arthritis of the hip in children²⁸), a more sensitive criterion is chosen. This relationship between the sensitivity and specificity of a diagnostic test can be portrayed with use of a *receiver operating characteristic (ROC) curve*. A receiver operating characteristic graph shows the relationship between the true-positive rate (sensitivity) on the y axis and the false-positive rate ($1 - \text{specificity}$) on the x axis plotted at each possible cutoff (Fig. 3). Overall diagnostic performance can be evaluated on the basis of the area under the receiver operating characteristic curve²⁹.

Measures of Effect

Measures of likelihood include probability and odds. *Probability* is a number, between 0 and 1, that indicates how likely an event is to occur on the basis of the number of events per the number of trials. The probability of heads on a coin toss is 0.5. *Odds* is the ratio of the probability of an event occurring to the probability of the event not occurring. The odds of heads coming up on a coin toss is 1 (0.5/0.5). Because probability and odds are related, they can be converted, where $\text{odds} = \text{probability}/(1 - \text{probability})$.

Relative risk (RR) can be determined in a prospective cohort study, where relative risk equals the incidence of disease in the exposed cohort divided by the incidence of disease in the nonexposed cohort (Table III). For example, if a prospective cohort study of skiers with deficiency of the anterior cruciate ligament shows a significantly higher proportion of subsequent knee injuries in skiers who are not treated with a brace (12.7%) than in those who are treated with a brace (2.0%), the risk ratio is 6.4 (12.7%/2.0%)³⁰. This can be interpreted as a 6.4 times higher risk of subsequent knee injury in a skier with anterior cruciate ligament deficiency who is not treated with a brace than in such a skier who is treated with a brace. A similar measurement in a retrospective case-control study (in which incidence cannot be determined) is the *odds ratio (OR)*, which is the ratio of the odds of having the disease in the study group to the odds of having the disease in the control group (Table III).

Factors that are likely to increase the incidence, prevalence, morbidity, or mortality of a disease are called risk factors. The effect of a factor that reduces the probability of an adverse outcome can be quantified by the relative risk reduction (RRR), the absolute risk reduction (ARR), and the number needed to treat (NNT) (Table III). The effect of a factor that increases the

probability of an adverse outcome can be quantified by the relative risk increase (RRI), the absolute risk increase (ARI), and the number needed to harm (NNH) (Table III).

Outcomes Assessment

Process refers to the medical care that a patient receives, whereas outcome refers to the result of that medical care. The emphasis of the outcomes assessment movement has been patient-derived outcomes assessment. Outcome measures include generic measures, condition-specific measures, and measures of patient satisfaction³¹. Generic measures, such as the Short Form-36 (SF-36), are used to assess health status or health-related quality of life, as based on the World Health Organization's multiple-domain definition of health^{32,33}. Condition-specific measures, such as the International Knee Documentation Committee (IKDC) knee score or the Constant shoulder score, are used to assess aspects of a specific condition or body system. Measures of patient satisfaction are used to assess various components of care and have diverse applications, including the evaluation of quality of care, health-care delivery, patient-centered models of care, and continuous quality improvement³⁴⁻³⁷.

The process of developing an outcomes instrument involves identifying the construct, devising items, scaling responses, selecting items, forming factors, and creating scales. A large number of outcomes instruments have been developed and used without formal psychometric assessment of their reliability, validity, and responsiveness to change. *Reliability* refers to the repeatability of an instrument. *Interobserver reliability* and *intraobserver reliability* refer to the repeatability of the instrument when used by different observers and by the same observer at different time-points, respectively. *Test-retest reliability* can be assessed by using the instrument to evaluate the same patient on two different occasions without an interval change in the patient's medical status. These results are usually reported with use of the *kappa statistic* or intraclass correlation coefficient. *Validity* refers to whether the instrument measures what it purports to measure. *Content validity* assesses whether an instrument is representative of the characteristic being measured according to expert consensus opinion (face validity). *Criterion validity* assesses an instrument's relationship to an accepted, "gold-standard" instrument. *Construct validity* assesses whether an instrument follows accepted hypotheses (constructs) and produces results consistent with theoretical expectations. Responsiveness to change assesses how an instrument's values change over the disease course and treatment.

TABLE IV Statistical Tests for Comparing Independent Groups and Paired Samples

Type of Data	Number of Groups	Independent Groups	Paired Samples
Continuous			
Normal	2	Student t test	Paired t test
Non-normal	2	Mann-Whitney U test	Wilcoxon signed-rank test
Normal	≥3	Analysis of variance	Repeated-measures analysis of variance
Non-normal	≥3	Kruskal-Wallis test	Friedman test
Ordinal			
2	2	Mann-Whitney U test	Wilcoxon signed-rank test
≥3	≥3	Kruskal-Wallis test	Friedman test
Nominal			
2	2	Fisher exact test	McNemar test
≥3	≥3	Pearson chi-square test	Cochran Q test
Survival	≥2	Log-rank test	Conditional logistic regression

Evidence-Based Medicine

Evidence-based medicine (EBM) involves the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients³⁸. Evidence-based medicine integrates the best research evidence with clinical expertise and patient values. The steps of evidence-based medicine involve converting the need for information into an answerable question; tracking down the best evidence to answer that question; critically appraising the evidence with regard to its validity, impact, and applicability; and integrating the critical

appraisal with clinical expertise and the patient's unique values and circumstances^{39,40}. The types of questions asked in evidence-based medicine are foreground questions pertaining to specific knowledge about managing patients who have a particular disorder. Evidence is graded on the basis of study design (see Table [Levels of Evidence for Primary Research Question] in Instructions to Authors of this issue of *The Journal*), with an emphasis on randomized clinical trials, and can be found in evidence-based databases (Evidence-Based Medicine Reviews [EBMR] from Ovid Technologies, the Cochrane Database of Systematic

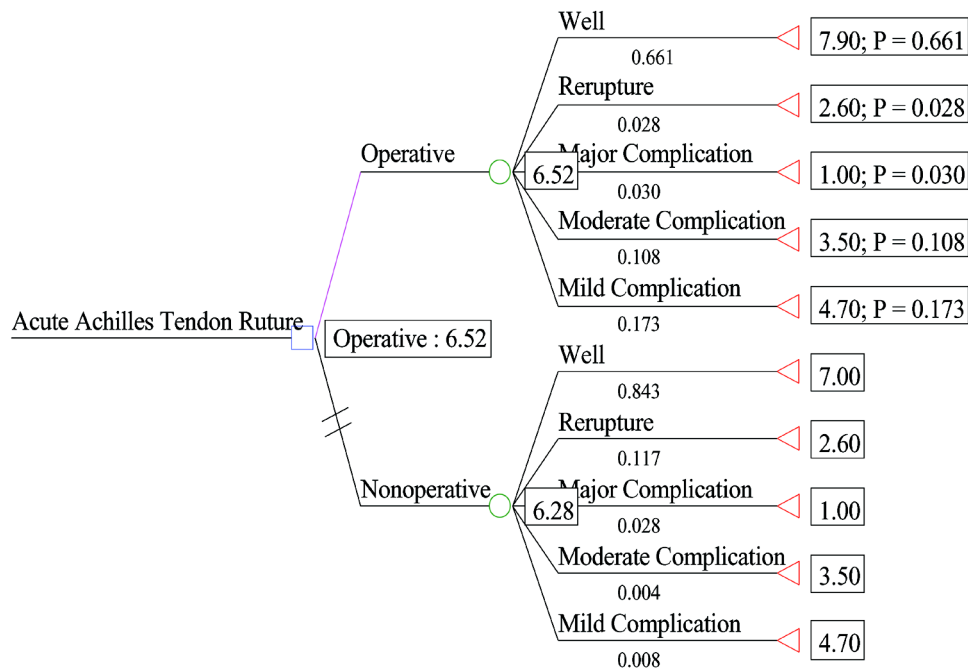


Fig. 4

Expected-value decision analysis tree for operative versus nonoperative management of acute Achilles tendon rupture⁵⁰. Decision nodes are represented by squares, chance nodes are represented by circles, and terminal nodes are represented by triangles. Mean outcome utility scores are listed to the right of the terminal node (0 to 10). Outcome probabilities are listed under the terminal node title (0 or 1). Operative treatment is favored because it has a higher expected value (6.52 compared with 6.28).

Reviews, Best Evidence, Clinical Evidence, National Guideline Clearinghouse, CancerNet, and MEDLINE) and evidence-based journals (*Evidence-Based Medicine* and *ACP Journal Club*).

A *systematic review* is a summary of the medical literature in which explicit methods are used to perform a thorough literature search and critical appraisal of studies. A specialized type of systematic review is *meta-analysis*, in which quantitative methods are used to combine the results of several independent studies (usually randomized clinical trials) to produce summary statistics. For example, a study that systematically reviews the literature (with criteria for inclusion and exclusion of studies) for reports comparing internal fixation and arthroplasty for the treatment of femoral neck fractures and then summarizes the outcomes and complications is considered a systematic review. On the other hand, a study in which the investigators systematically review the literature (with criteria for inclusion and exclusion of studies) and then combine the patient data to perform a new statistical analysis is considered a meta-analysis⁴¹. Clinical pathways or *clinical practice guidelines* (CPG) are algorithms that are developed, on the basis of the best available evidence, to standardize processes and optimize outcomes. They may also potentially reduce errors of omission and commission, reduce variations in practice patterns, and decrease costs.

Decision analysis is a methodological tool that allows quantitative evaluation of decision-making under conditions of uncertainty⁴²⁻⁴⁴. The rationale underlying explicit decision analysis is that a decision must be made, often under circumstances of uncertainty, and that rational decision theory optimizes expected value. The process of expected-value decision analysis involves the creation of a decision tree to structure the decision problem, determination of outcome probabilities and utilities (patient values), fold-back analysis to calculate the expected

value of each decision path to determine the optimal decision-making strategy (Fig. 4), and sensitivity analysis to determine the effect of varying outcome probabilities and utilities on decision-making (Fig. 5). Decision analysis can identify the optimal decision strategy and how this strategy changes with variations in outcome probabilities or patient values. This process, whether used explicitly or implicitly, integrates well with the newer doctor-patient model of shared decision-making.

Economic evaluative study designs in medicine include cost-identification studies, cost-effectiveness analysis, cost-benefit analysis, and cost-utility analysis^{45,46}. In *cost-identification studies*, the costs of providing the treatment are identified. In *cost-effectiveness analysis*, the costs and clinical outcome are assessed and reported as cost per clinical outcome. In *cost-benefit analysis*, both costs and benefits are measured in monetary units. In *cost-utility analysis*, cost and utility are measured and are reported as cost per quality-adjusted life-year (QALY).

Biostatistics

The scale on which a characteristic is measured has implications for the way in which information is summarized and analyzed. Data can be categorical, ordinal, or continuous. *Categorical data* indicate types or categories and can be thought of as counts. The categories do not represent an underlying order. Examples include gender and a dichotomous (yes/no, success/failure) outcome. Categorical data are also called nominal data. Categorical data generally are described in terms of proportions or percentages and are reported in tables or bar charts. If there is an inherent order among categories, then the data are *ordinal*. The numbers represent an order but are not necessarily to scale. Examples include cancer stages and injury grades. Ordinal data generally are also described in terms of proportions or percent-

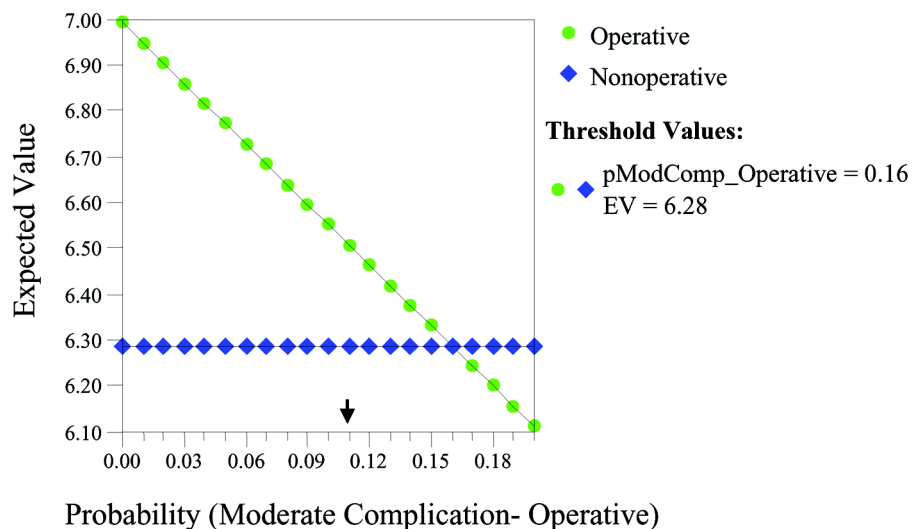


Fig. 5

Sensitivity analysis for operative versus nonoperative management of acute Achilles tendon rupture⁵⁰. The probability of a wound complication from operative treatment is varied on the x axis. The lines represent the expected value for the operative and nonoperative decisions. Above the threshold value (probability of wound complication from operative treatment = 16%), nonoperative treatment is favored.

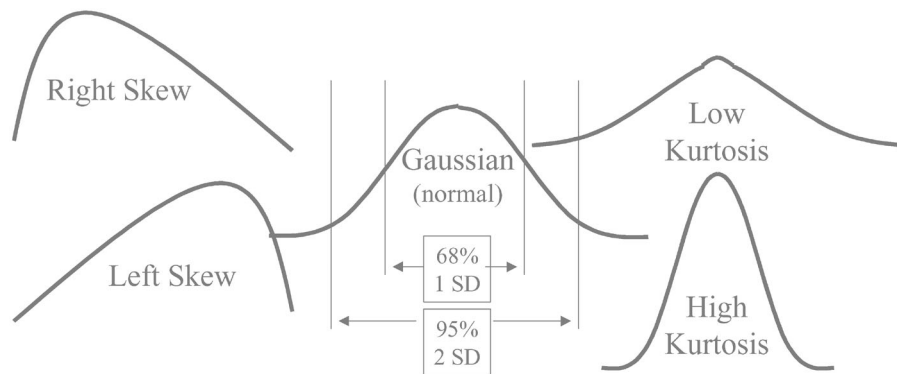


Fig. 6
Data distributions.

ages and are reported in tables or bar charts. *Continuous data* are observations on a continuum for which the differences between numbers have meaning on a numerical scale. Examples include age, weight, and distance. When a numerical observation can have only integer values, the scale of measurement is called *discrete*. Continuous data are generally described in terms of mean and standard deviation and can be reported in tables or graphs.

Data can be summarized in terms of measures of central tendency, such as *mean*, *median*, and *mode*, and in terms of measures of dispersion, such as range, *standard deviation*, and percentiles. Data can be characterized by different distributions, such as the normal (Gaussian) distribution, skewed dis-

tributions, and bimodal distributions (Fig. 6).

Univariate, or bivariate, analysis assesses the relationship of a single independent and a single dependent variable. Statistical tests for comparing means of continuous variables that are normally distributed include the *Student t test* for two independent groups and the *paired t test* for paired samples. For continuous or categorical variables that are not normally distributed, nonparametric statistical tests for comparing medians include the Mann-Whitney U test (also known as the *Wilcoxon rank-sum test*) to compare two independent groups and the *Wilcoxon signed-rank test* to compare paired samples (Table IV). *Analysis of variance (ANOVA)* is used to compare means of three or

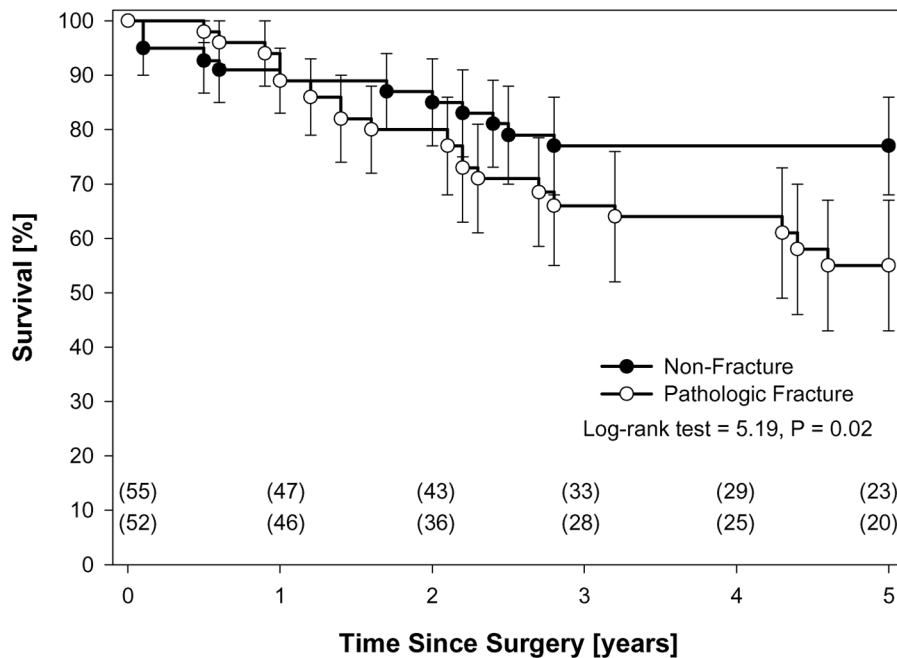


Fig. 7
Kaplan-Meier estimated survivorship curves comparing survival rates between patients who had osteosarcoma with a pathologic fracture and those who had osteosarcoma without a fracture⁵¹. The estimated rates were significantly lower for patients with a pathologic fracture (log-rank test = 5.19, $p = 0.02$). The error bars around the survivorship curves represent 95% confidence intervals derived by Greenwood's formula. The numbers of patients on whom the estimates were based are shown in parentheses.

more independent groups in which the data are normally distributed. The main result is the F test, which yields a p value indicating whether there is an overall significant difference. To determine if there are significant differences between individual groups, post-hoc tests are used to perform multiple pairwise comparisons between the groups; these tests include the Bonferroni, Tukey, Newman-Keuls, Scheffé, Fisher, and Dunnett procedures. The Kruskal-Wallis test is used to compare medians of three or more independent groups in situations where the data do not follow a normal distribution. The Kruskal-Wallis test is a nonparametric alternative to analysis of variance. Repeated-measures analysis of variance is used for normally distributed variables when the study has matched subjects. The nonparametric test to compare medians among three or more matched groups is called the Friedman test.

Statistical tests used to compare proportions for categorical or ordinal variables include the Pearson chi-square test for two or more independent groups and the Fisher exact test when expected cell frequencies are small (five or less). For matched samples, the McNemar test is used for two variables and the Cochran Q test is used for three or more variables (Table IV).

Statistical tests that are used to determine associations include the Pearson product-moment correlation (r) for normally distributed continuous variables, Spearman rank-order correlation (ρ) for nonparametric variables, and Kendall rank-correlation for ordinal variables.

Survivorship analysis is used to analyze data when the outcome of interest is the time until an event occurs. A group of patients is followed to determine if they experience the event of interest. The end point in survivorship analysis can be mortality or a clinical end point such as revision of a total joint replacement. A patient is censored when the event of interest does not occur to that individual during the study period. Survival is the length of time from a patient's entry into the study until the event of interest or until the time of censoring. The survival time for each patient is rarely known when a survival curve is constructed. Instead, the length of time that the patient's total joint replacement has survived so far or the length of time that it survived before it had to be revised is known. Often, patients have not had a failure at the end of the study period but remain at risk for failure in the future. These are examples in which information is censored because survival time is partially observed. Survivorship data are typically analyzed with use of the *Kaplan-Meier* product-limit method, in which the survivorship (freedom from the event) is calculated every time that an event occurs but not at censored times⁴⁷. *Kaplan-Meier* analysis is used when the actual date of the end point is known. End points that have not been reached are treated as censored at the date of the last follow-up for the analysis. Survivorship analysis produces a life table showing the number of failures occurring within time intervals and the number of patients withdrawn during the interval. A survivorship curve can be plotted to illustrate the percentage of patients free from failure (event-free) on the vertical axis and the follow-up time since the surgery on the horizontal axis (Fig. 7). The 95% confidence intervals can be constructed around the curve at selected time-points with use

of Greenwood's formula⁴⁸. Survivorship for different groups can be compared by the log-rank test for comparing the equality of the curves⁴⁹.

Multivariate analysis explores relationships between multiple variables. *Regression* is a method of obtaining a mathematical relationship between an outcome variable (Y) and an explanatory variable (X) or a set of independent variables (X_i 's). Linear regression is used when the outcome variable is continuous with the goal of finding the line that best predicts Y from X . Multiple regression fits data to a model that defines Y as a function of two or more explanatory variables or predictors. Logistic regression is used when the outcome variable is binary or dichotomous, and it has become the most common form of multivariate analysis for non-time-related outcomes. Other regression methods include time-to-event data (Cox proportional-hazards regression) and count data (Poisson regression). Regression modeling is commonly used to predict outcomes, or to establish independent associations (controlling for confounding and *collinearity*) among predictor or explanatory variables. For example, logistic regression can be used to determine predictors of septic arthritis versus transient synovitis of the hip in children on the basis of an array of presenting demographic, laboratory, and imaging variables²⁸. Similarly, linear regression can be used to determine independent determinants of patient outcome measured with use of a continuous outcome instrument³⁷. Because many variables usually influence a particular outcome, it is necessary to use multivariate analysis. The primary goal of multivariate analysis is to identify, from among the many patient and surgical variables observed and recorded, those most related to the outcome. Most multivariate analyses generate a tremendous amount of information, and proper interpretation requires expertise. It is an advantage to have a colleague trained in statistical methodology involved in the multivariate analysis.

Overview

Epidemiology and biostatistics are the essential tools of clinical research. An understanding of study design, hypothesis testing, diagnostic performance, measures of effect, outcomes assessment, evidence-based medicine, and biostatistics is essential both for the investigator conducting clinical research and for the practitioner interpreting clinical research reports.

NOTE: The authors thank Dr. James Wright and Dr. Robert Marx for their constructive comments.

Mininder S. Kocher, MD, MPH
David Zurakowski, PhD
Department of Orthopaedic Surgery, Children's Hospital, 300 Longwood Avenue, Boston, MA 02115. E-mail address for M.S. Kocher: mininder.kocher@childrens.harvard.edu

The authors did not receive grants or outside funding in support of their research or preparation of this manuscript. They did not receive payments or other benefits or a commitment or agreement to provide such benefits from a commercial entity. No commercial entity paid or directed, or agreed to pay or direct, any benefits to any research fund, foundation, educational institution, or other charitable or nonprofit organization with which the authors are affiliated or associated.

Glossary

- Absolute risk reduction (ARR):** Difference in risk of adverse outcomes between experimental and control participants in a trial.
- Alpha (type-I) error:** Error in hypothesis testing where a significant association is found when there is no true significant association (rejecting a true null hypothesis). The alpha level is the threshold of statistical significance established by the researcher ($p < 0.05$ by convention).
- Analysis of variance (ANOVA):** Statistical test to compare means among three or more groups (F test).
- Beta (type-II) error:** Error in hypothesis testing where no significant association is found when there is a true significant association (rejecting a true alternative hypothesis).
- Bias:** Systematic error in the design or conduct of a study. Bias threatens the validity of the study.
- Blinding:** Element of study design in which patients and/or investigators do not know who is in the treatment group and who is in the control group. The term *masking* is often used.
- Case-control study:** Retrospective observational study design that involves identifying cases with the outcome of interest and controls without the outcome and then looking back to see if they had the exposure of interest.
- Case series:** Retrospective observational study design that describes a series of patients with an outcome of interest or who have undergone a particular treatment. There is no control group.
- Categorical data:** Variable whose values are categories (nominal variable, qualitative data).
- Censored data:** In survivorship analysis, an observation whose outcome is unknown because the patient has not had the event of interest or is no longer being followed.
- Chi-square test:** Statistical test to compare proportions or categorical data between groups.
- Clinical practice guideline (CPG):** A systematically developed, evidence-based statement designed to standardize the process of care and optimize the outcome of care for specified clinical circumstances.
- Cohort study:** Prospective observational study design that involves the identification of a group or groups with the exposure or condition of interest and then follows the group or groups forward for the outcome of interest.
- Colinear:** In multivariate analysis, two or more independent variables that are not independent of each other.
- Conditional probability:** Probability of an event, given that another event has occurred.
- Confidence interval (CI):** Quantifies the precision of measurement. It is usually reported as the 95% confidence interval, which is the range of values within which there is a 95% probability that the true value lies.
- Confounder:** A variable that has independent associations with both the dependent and the independent variables, thus potentially distorting their relationship.
- Construct validity:** Psychometric property of an outcome instrument assessing whether the instrument follows accepted hypotheses (constructs).
- Content validity:** Psychometric property of an outcome instrument assessing whether the instrument is representative of the characteristic being measured (face validity).
- Continuous variable:** Variable whose values are numerical on a continuum scale of equal intervals and able to have fractions (interval, ratio, numerical, quantitative data).
- Controlling for:** Term used to describe when confounding variables are adjusted in the design or analysis of a study in order to minimize confounding.
- Correlation:** A measure of the relationship or strength of association between two variables.
- Cost-benefit analysis:** Economic evaluation of the financial costs compared with the benefits. Both are measured in monetary units. The result is reported as a ratio.
- Cost-effectiveness analysis:** Assesses the net costs and clinical outcome. The result is reported as a ratio of cost per clinical outcome.
- Cost-identification analysis:** Assesses only the net and component costs of an intervention. The result is reported in monetary units.
- Cost-utility analysis:** Assesses the net costs of the intervention and the patient-oriented utility of outcomes. The result frequently is reported as the cost per quality-adjusted life-year (QALY).
- Covariate:** An explanatory or confounding variable in a research study.
- Criterion validity:** Psychometric property of an outcome instrument assessing its relationship to an accepted, "gold-standard" instrument.
- Crossover study:** Prospective experimental study design that involves the allocation of two or more experimental treatments, one after the other, in a specified or random order to the same group of patients.
- Cross-sectional study:** Observational study design that assesses a defined population at a single point in time for both exposure and outcome (survey).
- Decision analysis:** Application of explicit, quantitative methods that analyze the probability and utility of outcomes in order to assess a decision under conditions of uncertainty.
- Dependent variable:** Outcome or response variable.
- Descriptive statistics:** Statistics, such as mean, standard deviation, proportion, and rate, used to describe a set of data.
- Discrete scale:** Scale used to measure variables that have integer values.
- Distribution:** Values and frequency of a variable (Gaussian, binomial, skewed).
- Effect size:** The magnitude of a difference considered to be clinically meaningful. It is used in power analysis to determine the required sample size.
- Evidence-based medicine (EBM):** Conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients.

Experimental study: Study design in which treatment is allocated (trial).

Factor analysis: Statistical method for analyzing relationships among a set of variables to determine underlying dimensions.

Failure: Generic term used for an event.

Fisher exact test: Statistical test used to compare proportions in studies with small sample sizes.

Hypothesis: A statement that will be accepted or rejected on the basis of the evidence in a study.

Incidence: Proportion of new cases of a specific condition in the population at risk during a specified time interval.

Independent events: Events whose occurrence has no effect on the probability of each other.

Independent variable: Variable associated with the outcome of interest that contributes information about the outcome in addition to that provided by other variables considered simultaneously.

Intention-to-treat analysis: Method of analysis in randomized clinical trials in which all patients randomly assigned to a treatment group are analyzed in that treatment group, whether or not they received that treatment or completed the study.

Interaction: Relationship between two independent variables such that they have a different effect on the dependent variable.

Internal consistency: Psychometric property of an outcome instrument regarding the degree to which individual items are related to each other.

Interobserver reliability: Reliability between measurements made by two observers.

Intraobserver reliability: Reliability between measurements made by one observer at two different points in time.

Kaplan-Meier method: Statistical method used in survivorship analysis to estimate survival rates at different times.

Kappa statistic: Statistic used to measure interobserver and intraobserver reliability.

Likelihood ratio (LR): Likelihood that a given test result would be expected in a patient with a condition compared with the likelihood in a patient without the condition. It is the ratio of the true-positive rate to the false-positive rate.

Log-rank test: Statistic used to compare two survival curves with censored observations.

Longitudinal study: Study in which the same patient is followed over multiple points in time.

Matching: Process of making two groups homogeneous for possible confounding factors.

Mean: Measure of central tendency. It is the sum of the values divided by the number in the sample.

Median: Measure of central tendency. It is the middle observation (50th percentile).

Meta-analysis: An evidence-based systematic review that uses quantitative methods to combine the results of several independent studies to produce summary statistics.

Mode: Measure of central tendency. It is the most frequent value.

Multiple comparisons: Pairwise group comparisons involving more than one p value.

Multivariate analysis: Analysis of a set of explanatory variables with respect to a single outcome or analysis of several outcome variables simultaneously with respect to explanatory variables.

Negative predictive value (NPV): Probability of not having the disease given a negative diagnostic test. It requires an estimate of prevalence.

Nominal data: Data that are classified into categories with no inherent order.

Nonparametric methods: Statistical tests making no assumption regarding the distribution of data.

Null hypothesis: Default testing hypothesis assuming no difference between groups.

Number needed to treat (NNT): Number of patients that must be treated in order to achieve one additional favorable outcome.

Observational study: Study design in which treatment is not allocated.

Odds: Probability that the event will occur divided by probability that the event will not occur.

Odds ratio (OR): Ratio of the odds of having a condition or outcome in the experimental group to the odds of having the condition or outcome in the control group (case-control study).

One-tailed test: Test in which the alternative hypothesis specifies a deviation from the null hypothesis in one direction only.

Ordinal variable: Variable that has an underlying order. The numbers used are not to scale.

Paired t test: Statistical test used to compare the difference or change in a continuous variable for paired samples.

Placebo: Inactive substance used to reduce bias by simulating the treatment under investigation.

Positive predictive value (PPV): Probability of having the disease given a positive diagnostic test. It requires an estimate of prevalence.

Power: Probability of finding a significant association when one truly exists (1 – probability of type-II [β] error). By convention, a power of $\geq 80\%$ is considered sufficient.

Prevalence: Proportion of individuals with a disease or characteristic in the study population of interest.

Probability: A number, between 0 and 1, indicating how likely an event is to occur.

Prospective study: Direction of inquiry is forward from the cohort. The events transpire after the study onset.

P value: Probability of a type-I (α) error. If the p value is small, it is unlikely that the results observed are due to chance.

Random sample: A sample of subjects from the population such that each has an equal chance of being selected.

Randomized clinical trial (RCT): Prospective experimental study design that randomly allocates eligible patients to the experimental or control group or to different treatment groups.

Receiver operating characteristic (ROC) curve: Graph showing the test's performance as the relationship between the true-positive rate and the false-positive rate.

Regression: Statistical technique for determining the relationship among a set of variables.

Relative risk (RR): Ratio of the incidence of the disease or outcome in the exposed cohort versus the incidence in the unexposed cohort (cohort study).

Relative risk reduction (RRR): Proportional reduction in adverse event rates between experimental and control groups in a trial.

Reliability: Measure of reproducibility of a measurement.

Retrospective study: The direction of inquiry is backward from the cases. The events transpired before the study onset.

Robust: A statistical method in which the test statistic is not affected by violation of underlying assumptions.

Sample: Subset of the population.

Selection bias: Systematic error in sampling the population.

Sensitivity: Proportion of patients who have the outcome who are classified as having a positive result.

Sensitivity analysis: Method in decision analysis used to determine how varying different components of a decision tree or model changes the conclusions.

Skewness: Statistical measure of the asymmetry of the distribution of values for a variable.

Specificity: Proportion of patients without the outcome who are classified as having a negative result.

Standard deviation: Descriptive statistic representing the deviation of individual values from the mean.

Student t test: Statistical test for comparison of means between two independent groups.

Survivorship analysis: Statistical method for analyzing time-to-event data.

Systematic review: Evidence-based summary of the medical literature that uses explicit methods to perform a thorough literature search and critical appraisal of studies.

Test-retest reliability: Psychometric property of the consistency of an instrument at different points in time without a change in status.

Two-tailed test: Test in which the alternative hypothesis specifies a deviation from the null hypothesis in either direction.

Univariate analysis: Analysis of the relationship of a single independent and a single dependent variable (bivariate analysis).

Utility: Measure of patient desirability or preference for various states of health and illness.

Validity: Degree to which a questionnaire or instrument measures what it is intended to measure.

Wilcoxon rank-sum test: Nonparametric version of the Student t test. It is also known as the Mann-Whitney U test.

Wilcoxon signed-rank test: Nonparametric version of the paired t test for comparing medians between matched groups.

References

- Hennekens CH, Buring JE. *Epidemiology in medicine*. Mayrent SL, editor. Boston: Little, Brown; 1987.
- Oxman AD, Sackett DL, Guyatt GH. Users' guides to the medical literature. I. How to get started. The Evidence-Based Medicine Working Group. *JAMA*. 1993;270:2093-5.
- Davidoff F, Haynes B, Sackett D, Smith R. Evidence based medicine. *BMJ*. 1995;310:1085-6.
- Sackett DL, Rosenberg WM. On the need for evidence-based medicine. *J Public Health Med*. 1995;17:330-4.
- Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. *BMJ*. 1996;312:71-2.
- Straus SE, Sackett DL. Using research findings in clinical practice. *BMJ*. 1998;317:339-42.
- Feinstein AR, Spitz H. The epidemiology of cancer therapy. I. Clinical problems of statistical surveys. *Arch Intern Med*. 1969;123:171-86.
- Feinstein AR, Pritchett JA, Schimpff CR. The epidemiology of cancer therapy. II. The clinical course: data, decisions, and temporal demarcations. *Arch Intern Med*. 1969;123:323-44.
- Feinstein AR, Pritchett JA, Schimpff CR. The epidemiology of cancer therapy. 3. The management of imperfect data. *Arch Intern Med*. 1969;123:448-61.
- Wright JG, Feinstein AR. A comparative contrast of clinimetric and psychometric methods for constructing indexes and rating scales. *J Clin Epidemiol*. 1992;45:1201-18.
- Wennberg J, Gittelsohn A. Small area variations in health care delivery. *Science*. 1973;182:1102-8.
- Wennberg J, Gittelsohn A. Variations in medical care among small areas. *Sci Am*. 1982;246:120-34.
- Wennberg JE. Dealing with medical practice variations: a proposal for action. *Health Aff (Millwood)*. 1984;3:6-32.
- Wennberg JE. Outcomes research: the art of making the right decision. *Internist*. 1990;31:26, 28.
- Wennberg JE. Practice variations: why all the fuss? *Internist*. 1985;26:6-8.
- Wennberg JE, Bunker JP, Barnes B. The need for assessing the outcome of common medical practices. *Annu Rev Public Health*. 1980;1:277-95.
- Chassin MR, Koseoff J, Park RE, Winslow CM, Kahn KL, Merrick NJ, Keesey J, Fink A, Solomon DH, Brook RH. Does inappropriate use explain geographic variations in the use of health care services? A study of three procedures. *JAMA*. 1987;258:2533-7.
- Kahn KL, Koseoff J, Chassin MR, Flynn MF, Fink A, Pattaphongse N, Solomon DH, Brook RH. Measuring the clinical appropriateness of the use of a procedure. Can we do it? *Med Care*. 1988;26:415-22.
- Park RE, Fink A, Brook RH, Chassin MR, Kahn KL, Merrick NJ, Koseoff J, Solomon DH. Physician ratings of appropriate indications for three procedures: theoretical indications vs indications used in practice. *Am J Public Health*. 1989;79:445-7.
- Millenson ML. *Demanding medical excellence: doctors and accountability in the information age*. Chicago: University of Chicago Press; 1997.
- Katz J. The Nuremberg Code and the Nuremberg Trial. A reappraisal. *JAMA*. 1996;276:1662-6.
- World Medical Association. Declaration of Helsinki. Recommendations guiding physicians in biomedical research involving human subjects. *JAMA*. 1997;277:925-6.
- Kocher MS, Mandiga R, Murphy JM, Goldmann D, Harper M, Sundel R, Ecklund K, Kasser JR. A clinical practice guideline for treatment of septic arthritis in children: efficacy in improving process of care and effect on outcome of septic arthritis of the hip. *J Bone Joint Surg Am*. 2003;85:994-9.
- Kocher MS, DiCanzio J, Zurakowski D, Micheli LJ. Diagnostic performance of clinical examination and selective magnetic resonance imaging in the evaluation of intraarticular knee disorders in children and adolescents. *Am J Sports Med*. 2001;29:292-6.
- Kocher MS. Ultrasonographic screening for developmental dysplasia of the hip: an epidemiologic analysis (Part I). *Am J Orthop*. 2000;29:929-33.
- Kocher MS. Ultrasonographic screening for developmental dysplasia of the hip: an epidemiologic analysis (Part II). *Am J Orthop*. 2001;30:19-24.
- Baron JA. Uncertainty in Bayes. *Med Decis Making*. 1994;14:46-51.

28. **Kocher MS, Zurakowski D, Kasser JR.** Differentiating between septic arthritis and transient synovitis of the hip in children: an evidence-based clinical prediction algorithm. *J Bone Joint Surg Am.* 1999;81:1662-70.
29. **Hanley JA, McNeil BJ.** The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology.* 1982;143:29-36.
30. **Kocher MS, Sterett WI, Briggs KK, Zurakowski D, Steadman JR.** Effect of functional bracing on subsequent knee injury in ACL-deficient professional skiers. *J Knee Surg.* 2003;16:87-92.
31. **Kane RL.** Outcome measures. In: Kane RL, editor. *Understanding health care outcomes research.* Gaithersburg, MD: Aspen; 1997. p 17-8.
32. **Patrick DL, Deyo RA.** Generic and disease-specific measures is assessing health status and quality of life. *Med Care.* 1989;27(3 Suppl):S217-32.
33. **Stewart AL, Ware JE Jr, editors.** *Measuring functioning and well-being: the medical outcomes study approach.* Durham: Duke University Press; 1992.
34. **Carr-Hill RA.** The measurement of patient satisfaction. *J Public Health Med.* 1992;14:236-49.
35. **Strasser S, Aharony L, Greenberger D.** The patient satisfaction process: moving toward a comprehensive model. *Med Care Rev.* 1993;50:219-48.
36. **Ware JE Jr, Davies-Avery A, Stewart AL.** The measurement and meaning of patient satisfaction. *Health Med Care Serv Rev.* 1978;1:1, 3-15.
37. **Kocher MS, Steadman JR, Briggs K, Zurakowski D, Sterett WI, Hawkins RJ.** Determinants of patient satisfaction with outcome after anterior cruciate ligament reconstruction. *J Bone Joint Surg Am.* 2002;84:1560-72.
38. **Sackett DL, Rosenberg WM.** The need for evidence-based medicine. *J R Soc Med.* 1995;88:620-4.
39. **Evidence-Based Medicine Working Group.** Evidence-based medicine. A new approach to teaching the practice of medicine. *JAMA.* 1992;268:2420-5.
40. **Sackett DL, Strauss SE, Richardson WS, Rosenberg W, Haynes RB.** *Evidence-based medicine: how to practice and teach evidence-based medicine.* 2nd ed. Edinburgh: Churchill Livingstone; 2000.
41. **Bhandari M, Devereaux PJ, Swiontkowski MF, Tornetta P 3rd, Obrebsky W, Koval KJ, Nork S, Sprague S, Schemitsch EH, Guyatt GH.** Internal fixation compared with arthroplasty for displaced fractures of the femoral neck. A meta-analysis. *J Bone Joint Surg Am.* 2003;85:1673-81.
42. **Birkmeyer JD, Welch HG.** A reader's guide to surgical decision analysis. *J Am Coll Surg.* 1997;184:589-95.
43. **Krahn MD, Naglie G, Naimark D, Redelmeier DA, Detsky AS.** Primer on medical decision analysis: part 4—analyzing the model and interpreting the results. *Med Decis Making.* 1997;17:142-51.
44. **Pauker SG, Kassirer JP.** Decision analysis. *N Engl J Med.* 1987;316:250-8.
45. **Detsky AS, Naglie IG.** A clinician's guide to cost-effectiveness analysis. *Ann Intern Med.* 1990;113:147-54.
46. **Weinstein MC, Stason WB.** Foundations of cost-effectiveness analysis for health and medical practices. *N Engl J Med.* 1977;296:716-21.
47. **Kaplan EL, Meier P.** Nonparametric estimation from incomplete observations. *J Am Statist Assoc.* 1958;53:457-81.
48. **Kalbfleisch JD, Prentice RL.** *The statistical analysis of failure time data.* New York: Wiley; 1980. p 10-4.
49. **Mantel N.** Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep.* 1966;50:163-70.
50. **Kocher MS, Bishop J, Marshall R, Briggs KK, Hawkins RJ.** Operative versus nonoperative management of acute Achilles tendon rupture: expected-value decision analysis. *Am J Sports Med.* 2002;30:783-90.
51. **Scully SP, Ghert MA, Zurakowski D, Thompson RC, Gebhardt MC.** Pathologic fracture in osteosarcoma: prognostic importance and treatment implications. *J Bone Joint Surg Am.* 2002;84:49-57.