

# CLINICAL EPIDEMIOLOGY AND BIOSTATISTICS: A PRIMER FOR ORTHOPAEDIC SURGEONS PART I

MININDER S. KOCHER, MD, MPH, AND DAVID ZURAKOWSKI, PhD

CHILDREN'S HOSPITAL ORTHOPAEDIC INSTITUTE FOR CLINICAL EFFECTIVENESS, HARVARD MEDICAL SCHOOL, HARVARD SCHOOL OF PUBLIC HEALTH, BOSTON, MA

## INTRODUCTION

Clinical epidemiology and biostatistics are the basic sciences of clinical research. This series of articles will provide a basic primer of clinical epidemiology and biostatistics for the orthopaedic surgeon.

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. Work by Wennberg and colleagues revealed large small-area variations in clinical practice, with some patients thirty times more likely to undergo an operative procedure than other patients with identical symptoms merely because of their geographic location<sup>1-6</sup>. Further critical research suggested that up to 40% of some surgical procedures might be inappropriate and that up to 85% of common medical treatments were not rigorously validated<sup>7-9</sup>. 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 1997<sup>10</sup>. Health maintenance organizations and managed care emerged. In addition, increasing federal, state, and consumer oversight were brought to bear on the practice of clinical medicine.

These forces have led to an increased focus on the clinical effectiveness of care. Clinical epidemiology provides the methodology to assess the clinical effectiveness of care. Part I of this series, presented here, provides an overview of the concepts of study design, hypothesis testing, measures of treatment effect, and diagnostic performance. Evidence-based medicine, outcomes assessment, data, and statistical analysis will be covered in Part II, to be published in next year's edition of *The Orthopaedic Journal at Harvard Medical School*. Examples from the orthopaedic literature and a glossary of terminology are provided.

**Dr. Kocher** is an Instructor in Orthopaedic Surgery, Harvard Medical School, and Director of the Program in Clinical Effectiveness, Harvard School of Public Health and Department of Orthopaedics, Children's Hospital, Boston MA

**Dr. Zurakowski** is the Principal Statistician in the Department of Orthopaedic Surgery, Children's Hospital, Boston MA

Address Correspondence To:

Mininder S. Kocher, M.D., M.P.H.  
Department of Orthopaedic Surgery  
Children's Hospital  
300 Longwood Avenue  
Boston, MA 02115

617.355.4849/ 617.739.3338 (fax)  
e-mail: mininder.kocher@tch.harvard.edu

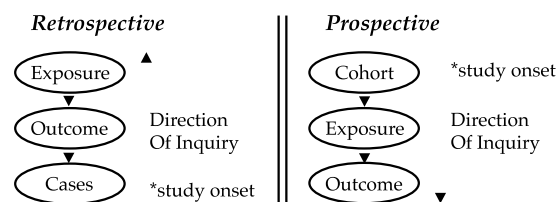


Figure 1: Prospective vs Retrospective Study Design

## 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 backwards from the cases and that the events of interest transpired before the onset of the study, or they 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.

All research studies are susceptible to invalid conclusions due to bias, confounding, and chance. *Bias* is the nonrandom 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, information, and nonresponder bias), and outcome determination (detection, recall, acceptability, and interviewer bias). A *confounder* is a variable having independent associations with both the *independent* (predictor) and *dependent* (outcome) variables, thus potentially distorting their relationship. Frequent confounders in clinical research include gender, age, socioeconomic status, and comorbidities. 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 selection, information, recall, and nonresponder bias. *Randomization* minimizes selection bias and equally distributes confounders. *Blinding* can decrease bias, and *matching* can decrease confounding. Confounders can sometimes be controlled *post hoc* with use of stratified analysis or multivariable methods. The effects of chance can be minimized by an adequate sample size based on *power* calcula-

Level	Levels of Evidence Description
1a	Systematic Review (with homogeneity) of Randomized Clinical Trials
1b	Individual Randomized Clinical Trial (with narrow confidence interval) Individual Inception Prospective Cohort Study with ≥80% follow-up
1c	All or None Case Series
2a	Systematic Review (with homogeneity) of Cohort Studies
2b	Individual Cohort Study Low-Quality Randomized Clinical Trial
2c	“Outcomes” Research Ecological Studies
3a	Systematic Review (with homogeneity) of Case-Control Studies
3b	Individual Case-Control Study
4	Case Series Low-Quality Cohort and Case-Control Studies
5	Expert Opinion
Grade	Grade of Recommendation Description
A	Consistent Level 1 Studies
B	Consistent Level 2 or 3 Studies; Extrapolations from Level 1 Studies
C	Level 4 Studies; Extrapolations from Level 2 or 3 Studies
D	Level 5 Studies; Troublingly Inconsistent or Inconclusive Studies of any Level

**Table I:** Oxford Centre for Evidence-Based Medicine Levels of Evidence\*  
\*<http://cebm.jr2.ox.ac.uk/docs/levels.html>

tions and appropriate *alpha levels*. The ability of study design to optimize validity while minimizing bias, confounding, and chance is recognized by the hierarchical levels of evidence and grades of recommendations established by the U.S. Preventative Services Task Force and the Oxford Centre for Evidence-Based Medicine on the basis of study design (Table I).

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 of 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 further studies but are not viewed as conclusive. A *case-control study* is a one in which the investigator identifies patients with an outcome of interest and controls without the outcome and then looks back retrospectively to identify possible causes or 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 very susceptible to multiple biases (particularly selection and detection bias). *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. 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 followed prospectively in order to determine outcomes and associations with risk factors. 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. The effects in a cohort study are frequently reported in terms of *relative risk*. Because these 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 long time-periods for completion, and are inefficient for the assessment of unusual outcomes or diseases.

Experimental trials may involve the use of concurrent controls, sequential controls (*cross-over trials*), or historical controls. The *randomized clinical trial (RCT)* with concurrent controls is the gold standard of clinical evidence as it provides the most valid conclusions (internal validity) by minimizing the effects of bias and confounding. A rigorous randomiza-

tion with enough patients is the best means of avoiding confounding. The performance of an RCT involves the construction of a protocol document that explicitly establishes eligibility criteria, sample size, informed consent, randomization, stopping rules, blinding, measurement, and data analysis. Because allocation is random, selection bias is minimized and confounders (known and unknown) are theoretically equally distributed between groups. *Blinding* minimizes performance, detection, interviewer, and acceptability bias. *Intention-to-treat analysis* minimizes non-responder 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 RCT is the epitome of clinical research designs, the disadvantages of RCTs include their expense, logistics, and time to completion. Accrual of patients and acceptance by clinicians may be problematic. With rapidly evolving technology, a new technique may become rapidly well accepted, making an existing RCT obsolete or a potential RCT difficult to accept. Ethically, RCTs require clinical equipoise (equality of treatment options in the clinician's judgment) for enrollment, interim stopping rules to avoid harm and evaluate adverse events, and truly informed consent. Finally, while RCTs have excellent internal validity, some have questioned their generalizability (external validity) because the practice pattern and the population of patients enrolled in an RCT 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.<sup>11,12</sup>

## 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 x 2 table of the possible outcomes of a study (Table II), we can see that the inference of a study is correct if a significant association is not found when there is no true associa-

Experiment	Truth	
	Not Significant	Significant
Not Significant	Correct	Type-II ( $\beta$ ) error
Significant	Type-I ( $\alpha$ ) error	Correct

Table II: Hypothesis Testing

P Value: probability of type-I ( $\alpha$ ) error  
 Power: 1 - probability of type-II ( $\beta$ ) error

tion 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* ( $\alpha$ ) 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* ( $\beta$ ) error wrongly concludes that there is no significant association (resulting in a “false negative” study that rejects a true alternative hypothesis).

The *P value* refers to the probability of the type-I ( $\alpha$ ) error. By convention, the alpha level of significance is set at 0.05, which means 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 from a statistical test, is a measure of the strength of evidence from 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 to reject it and conclude that the result is statistically 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 de-emphasize p values because a significance level of  $P < 0.05$  is arbitrary, a strict cutoff point can be misleading (there is little difference between  $P = 0.049$  and  $P = 0.051$ , yet 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 significant without 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 statistically significant or not, the 95% CI 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.

*Power* is the probability of finding a significant association if one truly exists and is defined as 1 - the probability of type-II ( $\beta$ ) error. By convention, acceptable power is set at  $\geq 80\%$ , which

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

**Table III:** Diagnostic Test Performance

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:	$[(\text{prevalence})(\text{sensitivity})]/[(\text{prevalence})(\text{sensitivity}) + (1-\text{prevalence})(1-\text{specificity})]$
Negative Predictive Value:	$[(1-\text{prevalence})(\text{specificity})]/[(1-\text{prevalence})(\text{specificity}) + (\text{prevalence})(1-\text{sensitivity})]$

means 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 ( $\alpha$ ) error as expressed by the p value. However, when a study demonstrates no significant association, the potential error of concern is the type-II ( $\beta$ ) 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 may have been too small. Thus, in a study that demonstrates no significant effect, the power of the study should be reported. The calculations for power analyses differ depending on the statistical methods utilized for analysis, however four elements are always involved in a power analysis:  $\alpha$ ,  $\beta$ , effect size, and sample size (n). Effect size is the difference that you want to be able to detect with the given  $\alpha$  and  $\beta$ . It is based on a clinical sense about how large a difference would be clinically meaningful. Low sample sizes, small effect sizes, and large variance decrease the power of a study. An understanding of power issues is important in clinical research to minimize resources when planning a study and to ensure the validity of a study. Sample size calculations are performed when planning a study. 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. Power analysis is performed after a study. Typically, alpha is set at 0.05, the sample size, effect size, and variance of the actual study are used, and the study's power is determined.

### DIAGNOSTIC PERFORMANCE

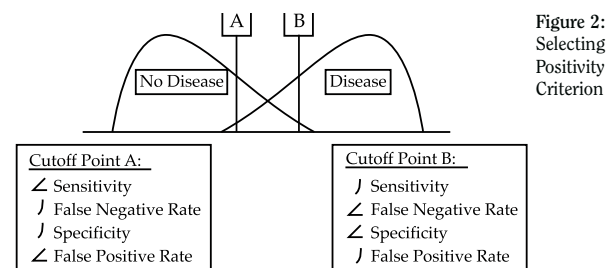
A diagnostic test can result in four possible scenarios: (1) *true positive* if the test is positive and the disease is present, (2) *false positive* if the test is positive and the disease is absent, (3) *true negative* if the test is negative and the disease is absent, and (4) *false negative* if the test is negative and the disease is present (Table III). The *sensitivity* of a test is the percentage (or

proportion) of patients who have the disease that are classified positive (true positive rate). A test with 97% sensitivity implies that of 100 patients with disease, ninety-seven will have a positive test. Sensitive tests have a low false-negative rate. A negative result on 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 negative (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 on 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 the gold standard of diagnosis in the same patients—for example, by comparing the results of magnetic resonance imaging with arthroscopic findings<sup>13</sup>.

Sensitivity and specificity are technical parameters of diagnostic testing performance and have important implications for screening and clinical practice guidelines<sup>14,15</sup>; however, they are less relevant in the typical clinical setting because the clinician does not know whether or not the patient has the disease. The clinically relevant questions are the probability that a patient has the disease given a positive result (*positive predictive value*) and the probability that a patient does not have the disease given a negative result (*negative predictive value*). The positive and negative predictive values are probabilities require an estimate of the prevalence of the disease in the population and can be calculated using equations that utilize Bayes' theorem.

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 one with 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 false-positive diagnosis of a condition

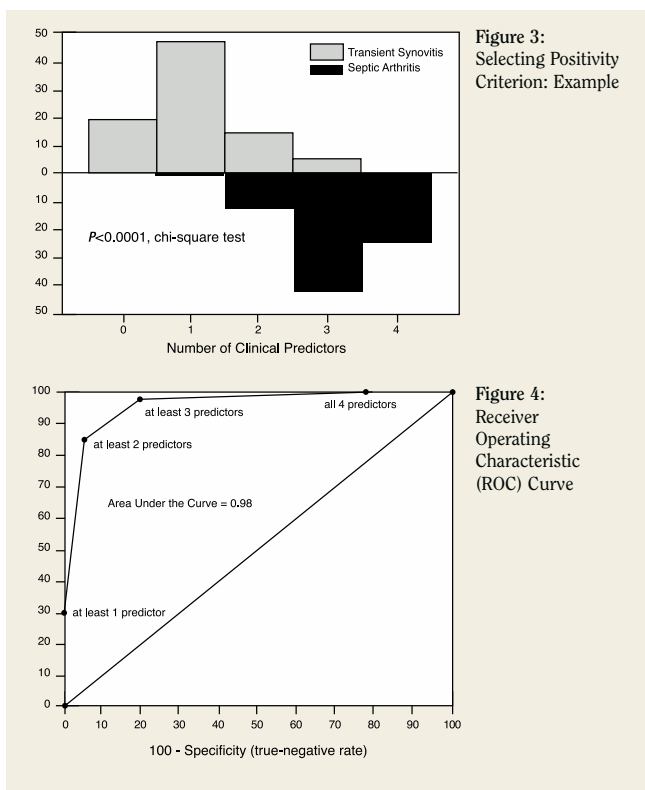


**Figure 2:** Selecting Positivity Criterion

	Adverse Events	No Adverse Events
Experimental Group	a	B
Control Group	c	d

**Table IV:** Treatment Effects

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):	$EER-CER$
Number Needed to Treat (NNT):	$1/ARR$



(such as septic arthritis of the hip in children<sup>16</sup>), a more sensitive criterion is chosen (Fig. 3). This relationship between the sensitivity and specificity of a diagnostic test can be portrayed with use of a *receiver operating characteristic (ROC) curve*. An ROC graph shows the relationship between the true positive rate (sensitivity) on the y-axis and the false positive rate (100-specificity) on the x-axis plotted at each possible cut-off (Figure 4). If a test discriminates well, its ROC curve approaches a true-positive rate of 100% and a true negative rate of 0%. On the other hand, a test that discriminates poorly has a diagonal ROC curve (45-degree line). Overall, the diagnostic performance can be evaluated by the area under the ROC curve. In the case of perfect discrimination the area under the curve will equal 1.0, while an area of 0.5 indicates random guessing.<sup>17</sup>

## 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 based on the number of events per the number of trials. The probability of heads on a coin toss is 0.5. *Odds* are the ratio of the probability of an event occurring to the probability of the event not occurring. The odds of flipping a heads on a coin toss is 1 (0.5/0.5). Because probability and odds are related, they can be converted where  $odds = probability/(1 - probability)$ .

*Relative risk (RR)* can be determined in a prospective cohort study, where RR equals the incidence of disease in the exposed cohort divided by the incidence of disease in the nonexposed cohort (Table IV). A similar measurement

in a retrospective case-control study (where incidence cannot be determined) is the *odds ratio (OR)*, which is the ratio of the odds of having the disease in the study group compared with the odds of having the disease in the control group (Table IV).

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 IV). 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 IV).

## 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).

**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. Threatens 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 which involves identifying cases with outcome of interest and controls without outcome, and then looking back to see if they had exposure of interest.

**Case Series:** Retrospective observational study design which describes a series of patients with an outcome of interest or who have undergone a particular treatment. No control group.

**Confidence Interval (CI):** Quantifies the precision of measurement. Usually reported as 95% CI, which is the range of values within which there is a 95% probability that the true value lies.

**Confounding:** A variable having independent associations with both the dependent and independent variables, thus potentially distorting their relationship.

**Cohort Study:** Prospective observational study design which involves the identification of group(s), with the exposure or condition of interest, and then following the group(s) forward for the outcome of interest.

**Controlling for:** Term used to describe when confounding variables are adjusted in the design or analysis of a study in order to minimize confounding.

**Crossover Study:** Prospective experimental study design which 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 which assesses a defined population at a single point in time for both exposure and outcome (survey).

**Dependent Variable:** Outcome or response variable.

**Distribution:** Values and frequency of a variable (Gaussian, binomial, skewed)

**Effect Size:** The magnitude of a difference considered to be clinically meaningful. Used in power analysis to determine the required sample size.

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

**Failure:** Generic term used for an event.

**Hypothesis:** A statement that will be accepted or rejected based on 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.

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

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

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

**Multiple Comparisons:** Pairwise group comparisons involving more than one *P*-value.

**Negative Predictive Value (NPV):** Probability of not having the disease given a negative diagnostic test. Requires an estimate of prevalence.

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

**Number Needed to Treat (NNT):** Number of patients needed to treat in order to achieve one additional favorable outcome.

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

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

**Odds Ratio:** Ratio of the odds of having condition/outcome in experimental group to the odds of having the condition/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.

**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. Requires an estimate of prevalence.

**Power:** Probability of finding a significant association when one truly exists (1-probability of type II ( $\beta$ ) error). By convention, power of 80% or greater 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 cohort. Events transpire after study onset.

**P-Value:** Probability of type I ( $\alpha$ ) error. If the P-value is small, then it is unlikely that the results observed are due to chance.

**Randomized Clinical Trial (RCT):** Prospective experimental study design which randomly allocates eligible patients to experimental vs control groups or different treatment groups.

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

**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 incidence of disease or outcome in exposed versus incidence in unexposed cohorts (cohort study).

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

**Retrospective Study:** Direction of inquiry is backwards from cases. Events transpired before study onset.

**Sample:** Subset of the population.

**Selection Bias:** Systematic error in sampling the population.

**Sensitivity:** Proportion of patients who have the outcome that are classified positive. **Sensitivity Analysis:** Method in decision analysis used to determine how varying different components of a decision tree or model change the conclusions.

**Specificity:** Proportion of patient without the outcome who are classified negative.

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

## References

1. Wennberg J, Gittelsohn A: Small area variations in health care delivery. *Science*, 182(117): 1102-8, 1973.
2. Wennberg J, Gittelsohn A.: Variations in medical care among small areas. *Sci Am*, 246(4): 120-34, 1982.
3. Wennberg JE: Dealing with medical practice variations: a proposal for action. *Health Aff (Millwood)*, 3(2): 6-32, 1984.
4. Wennberg JE: Outcomes research: the art of making the right decision. *Internist*, 31(7): 26, 28, 1990.
5. Wennberg JE: Practice variations: why all the fuss? *Internist*, 26(4): 6-8, 1985.
6. Wennberg JE, Bunker JP, Barnes B: The need for assessing the outcome of common medical practices. *Annu Rev Public Health*, 1: 277-95, 1980.
7. Chassin MR: Does inappropriate use explain geographic variations in the use of health care services? A study of three procedures [see comments]. *JAMA*, 258(18): 2533-7, 1987.
8. Kahn KL, Kosecoff J, Chassin M R, 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*, 26(4): 415-22, 1988.
9. Park RE, Fink A, Brook RH, Chassin MR, Kahn KL, Merrick NJ, Kosecoff J, Solomon DH: Physician ratings of appropriate indications for three procedures: theoretical indications vs indications used in practice. *Am J Public Health*, 79(4): 445-7, 1989.
10. Millenson ML. Demanding Medical Excellence. Chicago: University of Chicago Press, 1997.
11. Katz J. The Nuremberg Code and the Nuremberg Trial. *JAMA*, 276:1662-6, 1996.
12. World Medical Organization. Declaration of Helsinki: Recommendations guiding physicians in biomedical research involving human subjects. *JAMA*, 277:925-6, 1997.
13. Koehler MS, DiCanzio J, Zurakowski D, Micheli LJ. Diagnostic performance of clinical examination and selective magnetic resonance imaging in the evaluation of intra-articular knee disorders in children and adolescents. *Am J Sports Med*, 2001, 29(3): 292-296.
14. Koehler MS. Ultrasonographic screening for developmental dysplasia of the hip: An epidemiologic analysis. Part I. *Am J Orthop*, 2000, 29(12): 929-933.
15. Koehler MS. Ultrasonographic screening for developmental dysplasia of the hip: An epidemiologic analysis. Part II. *Am J Orthop*, 2001, 30(1):19-24.
16. Koehler 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*, 1999, 81A:1662-1670.
17. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143:29-36, 1982.