

Reading and Reviewing the Orthopaedic Literature: A Systematic, Evidence-based Medicine Approach

Kurt P. Spindler, MD, John E. Kuhn, MD, MS, Warren Dunn, MD, MPH,
Charles E. Matthews, PhD, Frank E. Harrell, Jr, PhD, and Robert S. Dittus, MD, MPH

Abstract

The principles of evidence-based medicine are rapidly gaining acceptance in the field of orthopaedic surgery. This approach to patient care requires a careful, systematic review of the literature to appropriately value the merit of studies. Systematic review assists the orthopaedic surgeon in interpreting study results and in understanding the relative validity of these results in the hierarchy of evidence. Sufficiently valid evidence-based information ultimately will help in making decisions regarding patient care.

J Am Acad Orthop Surg 2005;13:220-229

Traditionally, dogma has ruled the education of physicians: one need only recall the authoritative professor who by his name or her stature alone influenced how we evaluate and examine our patients. Despite the view that we rely on science to guide our approach to patients and to treatment decision-making, physicians frequently change their practice based on the opinions of charismatic and dogmatic (albeit usually experienced) authority figures. Although often correct, these changes in the practice of orthopaedics likely are influenced more by the volume of the authority's voice and the persuasiveness of his or her tone than by the scientific validity of the message.

During the last decade, a quiet movement has rapidly been gaining momentum that undoubtedly will change the way orthopaedics will come to be practiced. This movement is evidence-based medicine.^{1,2} The message of evidence-based medicine is not complicated. It is essentially a scientific approach by which the clinician collects and interprets informa-

tion about medical treatments, procedures, devices, diagnoses, and prognoses and applies this information to his or her practice. This approach attempts to remove as much opinion and bias as possible and to ensure that changes in the practice of orthopaedics are driven by the best science available.³

Evidence-based Medicine

The most frequently quoted definition of evidence-based medicine is that it is "the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients."¹ It requires us to make decisions by critically reading and reviewing the literature, then weighing the findings reported in studies by the scientific validity of the work and the researchers' approach. Evidence-based medicine asks us to be more critical about the changes that we make in our practice. It requires us to use the best evidence by placing more value on well-

designed and well-executed clinical investigations and less value on expert opinion and uncontrolled obser-

Dr. Spindler is Professor and Vice Chairman, Department of Orthopaedics and Rehabilitation, and Chief, Division of Sports Medicine, Vanderbilt University Medical School, Nashville, TN. Dr. Kuhn is Associate Professor and Chief, Shoulder Surgery, Vanderbilt University Medical School. Dr. Dunn is Assistant Professor, Department of Orthopaedics and Rehabilitation, and Assistant Professor, General Internal Medicine and Public Health, Health Services Research Center, Vanderbilt University Medical School. Dr. Matthews is Assistant Professor, General Internal Medicine and Public Health, Health Services Research Center, Vanderbilt University Medical School. Dr. Harrell is Professor and Chairman, Department of Biostatistics, Vanderbilt University Medical School. Dr. Dittus is the Harvie Branscomb Distinguished Professor, Albert and Bernard Werthan Professor of Medicine, Chief, Division of General Internal Medicine, and Director, Center for Health Services Research, Vanderbilt University Medical School.

Dr. Spindler or the department with which he is affiliated has received research or institutional support from Aircast and from Smith & Nephew. None of the following authors or the departments with which they are affiliated has received anything of value from or owns stock in a commercial company or institution related directly or indirectly to the subject of this article: Dr. Kuhn, Dr. Dunn, Dr. Matthews, Dr. Harrell, and Dr. Dittus.

Reprint requests: Dr. Kuhn, Vanderbilt Sports Medicine, Suite 4200, Medical Center East, South Tower, 1215 21st Avenue South, Nashville, TN, 37232-8774.

Copyright 2005 by the American Academy of Orthopaedic Surgeons.

Table 1
The Five Steps of the
Evidence-based Medicine
Approach³

1. Formulate answerable questions
2. Gather the evidence
3. Appraise the evidence
4. Implement the evidence
5. Evaluate the process

vational studies (eg, case reports and case series).

Five fundamental steps characterize the use of an evidence-based medicine approach (Table 1). Bernstein³ has reviewed the concepts of evidence-based medicine for the orthopaedic audience and provides a thorough review of these fundamental steps. Each step is important. The following discussion, however, focuses on the third step, appraising the evidence.

Study Types, Study Designs, Levels of Evidence

The different types of clinical investigation may be classified into broad types based on the purpose of the research² (Table 2). Clinical investigation also may be classified by the types of study design; these include

case reports, case series, ecologic studies, cross-sectional studies, case-control studies, cohort studies, experimental clinical trials, and meta-analyses. Each type of research, characterized by its purpose, determines the appropriateness of the study design.^{2,4,5} Also, weaker study designs are frequently used to generate hypotheses in a field, whereas stronger designs are used to test hypotheses. Both hypothesis-generation and hypothesis-testing studies are critical to the advancement of clinical medicine.

Types of Studies

Therapy research is designed to test the efficacy (how well an intervention works under ideal conditions) and effectiveness (how well an intervention works in clinical practice) of drug treatments, surgical treatments, or other interventions. An example is the use of minimally invasive surgical approaches to joint arthroplasty. The randomized controlled clinical trial is the preferred study design for examining the efficacy or effectiveness of a therapy.

Diagnosis research is intended to demonstrate the validity and reliability of new diagnostic tests. An example would be a new physical examination test to identify lesions of the superior labrum of the shoulder. Prospective cohort studies are the pre-

ferred study design for determining the reliability and validity of new diagnostic tests; clinical trials are preferred for examining their merit in a strategy of care.

Screening research is a type of diagnosis research and is conducted to determine the value of a screening test designed to detect a disease or condition at an early stage. An example is a serologic test to identify people at risk for osteoporosis. The same research design preferences hold for screening research as for diagnosis research: prospective cohort studies and clinical trials.

Prognosis research is conducted to describe the natural history of an illness or clinical condition or the clinical outcomes following a strategy of care. An example would be research to determine whether small rotator cuff tears progress to larger tears. For this type of research, a prospective cohort study is the preferred design. Historical cohort studies sometimes may be used effectively to examine questions of prognosis.

Finally, risk factor research is undertaken to determine whether a potential anatomic feature or exposure to a particular agent may predispose a person to the development of a condition or a disease. An example of this type of research would be to determine whether variations in acromion morphology cause rotator cuff dis-

Table 2
Types of Research Studies and the Preferred Study Design

Type of Research	Purpose	Preferred Study Design
Therapy	Tests efficacy or effectiveness of new treatment by surgery	Randomized controlled trial
Diagnosis	Determines the reliability and validity of a new diagnostic test or examination finding	Randomized controlled trial Prospective cohort study
Screening	Tests the value of a screening diagnostic test in the general population or in a defined subgroup	Randomized controlled trial Prospective cohort study
Prognosis	Determines the outcomes of a disease in the general population or in a defined subgroup	Prospective cohort study
Risk factor	Determines whether a particular risk factor is related to the development of a disease	Prospective cohort study Historical cohort study Case-control study

ease. Prospective cohort studies provide the strongest evidence but are often impractical because of the required study sample size, study cost, or necessary length of time for follow-up. These studies also may be impossible to implement; for example, it is unethical to follow persons with a known illness without intervention. Therefore, historical cohort studies and case-control studies are commonly used in risk factor research.

Study Designs

As mentioned, each of these types of research, characterized by its purpose, is associated with a preferred study design (or designs) that either will provide the strongest evidence or will be the most practical to implement. For example, therapy research is best studied with a randomized controlled trial. Hallmarks of the randomized controlled trial are that subjects are randomly assigned to specific treatment groups and are followed prospectively for the outcome of interest. Randomization attempts to ensure that the treatment groups under evaluation are initially similar with respect to known (measurable) and unknown (immeasurable) sources of bias. Randomized controlled trials permit specific examination of the effect of carefully specified treatments on specific measurable outcomes. The prospective element of the design clearly establishes the temporal sequence of the treatment on the outcome for the study.

Prospective designs, which are so useful in therapy research, also are preferred for testing hypotheses related to diagnosis research and screening research. However, cross-sectional studies are commonly used to describe the prevalence of a disease (critical in understanding the ultimate predictive value of a diagnostic test), the prevalence of a human anatomic feature, or the clinical utilization of a test or treatment. In addition, cross-

sectional designs are important for generating hypotheses related to the potential merit of diagnostic or screening tests. In a cross-sectional design, data are collected simultaneously from a population of interest. In that population, some patients will have the condition and others will not. Within these two populations, researchers could determine, for example, whether a new physical examination test is associated with superior labral tears, or whether a new serology test is associated with osteoporosis at an early stage. Studies examining the merit of new diagnostic or screening tests compare the new test either to a "gold standard" test or to its accuracy in determining the actual clinical outcome of interest. Because current gold standard tests are often imperfect, careful interpretation is required when comparing a new test to a gold standard test. There are many reasons why a diagnostic test might be recommended in actual clinical practice; thus, each test's performance must be carefully understood. The effectiveness of the new test should at least be equal to, if not superior to, currently used tests before it is recommended.

The prospective cohort study is the preferred design for prognosis research. In this design, a study population (cohort) is followed systematically over time. To examine the effect of a specific exposure or patient characteristic on the outcome, various statistical approaches may be taken. These exposures or characteristics may be either protective against the outcome of interest or associated with an increased incidence of the outcome. Such factors can take various forms, such as being dichotomous (eg, presence or absence) or continuous (eg, age). Prospective cohort studies can determine the natural history of a condition. For example, a population of patients with small rotator cuff tears could be followed to see whether the small tears enlarge after falls on the outstretched arm.

Finally, risk factor research should be done using either a prospective or historical cohort or a case-control study design. In contrast with randomized prospective clinical trials, cohort studies may be more prone to bias because only measurable confounding factors can be controlled statistically, and appropriate comparison groups may not be readily available. The prospective design for testing a new treatment permits researchers to determine the research question and the outcomes of interest before the experiment starts and before data are collected. The prospective design also ensures that the data collected are appropriate to answer the research question. By contrast, in a historical cohort or a case-control study, the research question is asked after data have been collected. One problem with these types of design is the limitation of the collected data, which often are not the most appropriate to answer the research question.

In a case-control study, patients with new diagnoses of a specific condition are compared to a control group without the disease. Data are then collected on previous exposure to a protective factor or a risk factor to determine whether cases and controls differ in their rates of exposure. For example, a study on incidence of rotator cuff tear with overhead assembly-line work as a risk factor would be such a source population study. Data are collected on past exposures (overhead assembly-line work) to determine whether cases and controls differ in their rates of exposure. A well-conducted case-control study provides an odds ratio describing the relationship between the exposure and the outcome. This ratio, with certain assumptions, may be considered a reasonable estimate of the true relative risk of the exposure for the outcome. However, this study design is susceptible to a large number of biases that must be carefully considered when both conducting

and interpreting the study. Examples of such biases include the process of selecting control subjects (a selection bias) and gathering exposure information from cases through personal interview (a type of measurement bias often referred to as recall bias). Another type of measurement bias, detection bias, can occur when the disease process influences the frequency of exposure measurement or the vigilance of a study monitor.

In conducting a cross-sectional, historical cohort, or case-control study, the difference between prevalent and incident cases must be addressed.⁶ Prevalent cases reflect the population that has the disease at a specific calendar time (eg, annual prevalence); prevalence includes cases existing in the population as well as new cases that emerge during the measurement interval. Because prevalent cases include both those persons who have had the disease for some time and those who have recently acquired the disease, associated exposure factors include true risk factors for developing the disease as well as factors that might prolong survival with the disease (actual protective factors). In contrast, incident cases reflect new cases of the disease that develop in the population over a specified duration. Factors associated with incident cases reflect only risk for developing the disease or the outcome of interest.

Cross-sectional studies for risk factors have additional concerns. One of these is the time relationship between exposure and disease. In a cross-sectional study, it is not possible to determine whether an exposure occurred before or after onset of the disease. Thus, a positive association between an exposure and a disease could mean the exposure causes the disease, but it could also mean that the disease causes the exposure. For example, in cross-sectional studies of cadaver shoulders, the hooked acromion does indeed seem to be associated with rotator cuff tears.⁷ How-

ever, it is unclear whether the hooked acromion caused the rotator cuff tear or whether the tear caused the hooked acromion. Cohort and case-control studies are thus stronger designs for determining causality (eg, find a group of patients with no cuff disease, assess their acromion morphology, and follow them to see whether those with a hooked acromion develop rotator cuff tears).

Study Design and the Hierarchy of Evidence

Each type of study design is appropriate in different situations depending on the nature of the question being posed. With respect to the efficacy or effectiveness of a therapeutic intervention, each design may be ordered as to the strength of evidence provided⁸ (Table 3). For example, a well-designed and executed double-blind randomized controlled prospective clinical trial with excellent follow-up provides stronger evidence for the use of diagnostics and therapeutics than do weaker designs. Although case reports and case series still have value (such as alerting clinicians to new diseases or alerting researchers to new treatments that may be worthy of study), study designs that use appropriate comparison groups and pay careful attention to sources of bias should be held in higher regard when accumulating evidence to change the way we practice.

This concept of ranking research studies in terms of their methodological strength is called the hierarchy of evidence. It is being used by many journals, including *The Journal of Bone and Joint Surgery*, *Clinical Orthopaedics and Related Research*, *Arthroscopy*, and *The American Journal of Sports Medicine* to classify published manuscripts. The hierarchy of evidence used by these journals and adopted by the AAOS Evidence-Based Practice Committee ([http://www.aaos.org/wordhtml/research/comitee/evidence/ebpc.](http://www.aaos.org/wordhtml/research/comitee/evidence/ebpc.htm)

htm) is presented in Table 3. In considering whether to change one's practice based on the results of an evidence-based study, it is imperative to know the type of study used in order to judge the methodologic strength of the study.

The Research Question or Hypothesis

Clinical studies are guided by a research question or hypothesis. A particular observational study will try to answer a research question, such as "What is the 30-day mortality associated with total knee replacement?" However, a similar study might also test a hypothesis, such as the following: "We hypothesize that older age is associated with a higher 30-day mortality following knee replacement, after adjusting for comorbidity." Other observational studies will be guided by a research question and will be descriptive, or "hypothesis generating." In hypothesis-generating studies, however, the strength of the evidence to be collected, and the type of study design used, will not be able to examine a specific hypothesis. Instead, ideas for hypotheses to be tested will be suggested through the descriptive results. Certain observational studies, and all experimental studies, are hypothesis testing and thus are guided by a primary hypothesis, and possibly by one or more secondary hypotheses.

The hypothesis is a carefully structured summary statement of the research being addressed, the population being studied, the nature of the comparison group, the metric being used to determine the result, and the magnitude and direction of the anticipated change between the study and comparison groups. The research question or hypothesis should be clearly stated in the introduction or in the methods section of the manuscript. The research question or hypothesis will determine the appropriate study design and analysis.

Table 3
Levels of Evidence for Primary Research Question*

Level	Type of Study			
	Therapeutic Studies— Investigating the results of treatment	Prognostic Studies— Investigating the effect of a patient characteristic on the outcome of disease	Diagnostic Studies— Investigating a diagnostic test	Economic and Decision Analyses—Developing an economic or decision model
I	High-quality RCT with statistically significant difference or no statistically significant difference but narrow confidence intervals Systematic review [†] of level I RCTs (and study results were homogeneous [‡])	High-quality prospective study [§] (all patients were enrolled at the same point in their disease with ≥80% follow-up of enrolled patients) Systematic review [†] of level I studies	Testing of previously developed diagnostic criteria on consecutive patients (with universally applied reference “gold standard”) Systematic review [†] of level I studies	Sensible costs and alternatives; values obtained from many studies; with multiway sensitivity analyses Systematic review [†] of level I studies
II	Lesser quality RCT (eg, <80% follow-up, no blinding, or improper randomization) Prospective [§] comparative study [¶] Systematic review [†] of level II studies or level I studies with inconsistent results	Retrospective [#] study Untreated controls from an RCT Lesser quality prospective study (eg, patients enrolled at different points in their disease or <80% follow-up) Systematic review [†] of level II studies	Development of diagnostic criteria on consecutive patients (with universally applied reference “gold standard”) Systematic review [†] of level II studies	Sensible costs and alternatives; values obtained from limited studies; with multiway sensitivity analyses Systematic review [†] of level II studies
III	Case-control study ^{**} Retrospective [#] comparative study [¶] Systematic review [†] of level III studies	Case-control study ^{**}	Study of nonconsecutive patients (without consistently applied reference “gold standard”) Systematic review [†] of level III studies	Analyses based on limited alternatives and costs; and poor estimates Systematic review [†] of level III studies
IV	Case series ^{††}	Case series	Case-control study Poor reference standard	Analyses with no sensitivity analyses
V	Expert opinion	Expert opinion	Expert opinion	Expert opinion

RCT = randomized clinical trial

*A complete assessment of quality of individual studies requires critical appraisal of all aspects of the study design

[†]A combination of results from two or more prior studies

[‡]Studies provided consistent results

[§]Study was started before the first patient enrolled

[¶]Patients treated one way (eg, cemented hip arthroplasty) compared with a group of patients treated in another way (eg, uncemented hip arthroplasty) at the same institution

[#]The study was started after the first patient enrolled

^{**}Patients identified for the study based on their outcome, called “cases” (eg, failed total arthroplasty), are compared to those who did not have that outcome, called “controls” (eg, successful total hip arthroplasty)

^{††}Patients treated one way with no comparison group of patients treated in another way

Data for this table are from <http://www.ejbs.org/misc/public/instrux.shtml> and http://www.cebm.net/levels_of_evidence.asp.

Bias

In any clinical investigation, relationships between exposure factors of interest and the clinical outcomes of interest ultimately will either prove to be real or will be mistaken as a result of bias, confounding factors, or sta-

tistical chance. No honest researcher intentionally introduces bias into a research effort. Nevertheless, like a bad ingredient unknowingly used by a good chef, bias can be a hidden poison in medical information, threatening the internal validity of a study. Bias can be difficult to detect, and the

clinician-reviewer must be ever vigilant to consider it. There are many (well over 30) types of bias that have been named and described. The user of clinical research must remain aware of them in order to best assess the merit of a clinical study. These biases are generally categorized into

two groups: selection bias and measurement bias. Confounding is sometimes referred to as a bias and given its own category.

Selection Bias

Selection bias may be introduced when the populations under study are first assembled. It is critically important to be sure that the two populations are similar. When they are different, the researcher cannot be certain that the difference in the effect of an intervention is the result of the intervention or the result of different characteristics of the two groups.

Selection bias can take many forms. In studying exposures to risk factors that may influence a disease, selection bias occurs when the study subjects have different probabilities of exposure to that risk factor. Therefore, to reduce the likelihood of selection bias, prospective studies should have well-defined entry criteria and recruitment processes. Clinical trials are best performed by randomly assigning patients to treatment groups. There are methods of randomizing patients (eg, simple, block, randomly varying block) that reduce the risk of bias. There are also pseudo-random methods of assigning patients that may be subject to bias (eg, by surgeon, by the days of the week that patients present to the clinic).

Even randomized trials may be at risk for a group assignment bias if clinicians are able to predict the next group assignment in a clinical trial (and clinicians have been shown to be notorious for attempting to break the randomization code). In such situations, a clinician's own bias toward a treatment steers patients into one group or the other. For example, with such a pseudo-randomized approach, it would not be difficult for the nurse to schedule a friend into a preferred group. The use of randomly varying small-block randomization schemes is one highly effective approach.

Another form of selection bias occurs as a result of patient follow-up.

In this form of bias, the study group available for analysis is determined by the group remaining in the study for which outcome measures can be performed. Thus, if there is a systematic reason why certain patients are likely to drop out of the study that might be related to the exposure or outcome variables, then selection bias is possible. However, if follow-up is "nondifferential"—that is, not related to the exposure or outcome of interest—then an important selection bias might not be present. This form of follow-up bias is categorized as a selection bias because it influences the population available for analysis, just as would a bias in assembling the study population from the beginning. Therefore, any prospective study should devote the necessary resources to achieve the highest possible follow-up (*The Journal of Bone and Joint Surgery* requires at least 80% for a level I study) with adequate duration so that the effect of the intervention is apparent and complete.

Measurement Bias

Measurement bias, sometimes referred to as misclassification bias, results from inaccuracies in the measurement of the exposure variables, outcome variables, or potential confounding variables. One form of measurement bias is detection bias, which happens when the outcome evaluations for the groups of patients are different. This difference might occur because one group has had more frequent testing or been tested using different instruments. Another form of measurement bias results from the use of exposure or outcome measurements that lack reliability or validity. Recent efforts in many fields of orthopaedics have produced outcome-assessment tools that have undergone rigorous testing to be certain they test what they are designed to test (a process called validation) and can consistently reproduce the same results (reliability). The use of valid and reliable outcome-assessment tools in-

creases the strength of a research project and allows for comparisons of different studies investigating similar questions.

Confounding

Confounding is sometimes given its own category as one of the reasons or variables that could affect an exposure-outcome relationship. Confounding occurs when some factor is statistically associated with both the exposure and outcome. Confounding factors can cause an apparent relationship where none truly exists, or they can mask a true relationship. For example, concomitant treatments other than the intervention targeted for study could be performed on the different exposure groups. Potential differences in outcomes could be a result of these concomitant treatments rather than of the studied intervention. Thus, a study hypothesis might be that meniscus allograft is effective in the relief of pain and improved patient outcomes. However, the study subjects who received meniscus allograft differ from those who did not receive allografts with respect to other procedures, such as osteotomy or articular cartilage restorative procedures. Therefore, it is difficult to know whether the outcomes seen are related to the meniscus allograft or to the other interventions. All interventions thus must be documented and controlled for to determine whether the outcomes are the result of the target intervention.

Controlling for Chance: Statistical Concepts of Importance

Selection and measurement biases must be managed during the design and execution phases of a clinical investigation. Confounding also should be addressed in the design and execution phases, but in addition it must be examined during analysis. Examining the role that chance might play

in explaining the results of a clinical study occurs during the analysis; however, the potential role of chance is taken into account during the design phase.

Every clinical investigator must become familiar with P values and study power.^{9,10} In traditional statistics, clinicians initially assume the hypothesis of no effect (the null hypothesis), then determine whether the data appear to be valid were the null hypothesis true. The P value is the probability of observing results as extreme as, or more extreme than, the observed findings if the null hypothesis were true. When the P value is sufficiently small (eg, $P < 0.001$ or $P = 0.05$), the null hypothesis is rejected.

Assessing evidence in this way is like being a juror in a trial. A suspect is accused of a crime (his guilt is the hypothesis). However, the suspect is presumed innocent until proven guilty (the null hypothesis, which evidence must refute). There must be sufficient evidence for the jury to find the suspect guilty beyond a reasonable doubt.¹¹ The P value is the probability of a false conviction (rejection of innocence).

Four potential outcomes exist in a jury trial. (1) An innocent individual may be acquitted (correct decision). (2) A guilty individual may be found guilty (correct decision). (3) An innocent individual may be found guilty (incorrect decision). (4) A guilty individual may be found innocent (incorrect decision). In essence, the P value is the probability of having made mistake no. 3, that an innocent person is found guilty (that is, that a difference exists between two treatments when in fact there is no difference). This is known as a type I error.

Following this analogy, a type II error is the probability of making mistake no. 4 (that is, that no difference exists between two treatments when in fact there is a difference). The power of a test is the probability of detecting a significant effect (1 minus the probability of making a type II error).

If the jury does not have sufficient evidence (or the study a sufficient number of subjects), then the jury may have to let a guilty individual go free (and the study must state that no difference exists between two treatments when in fact a difference does exist).

Another point worthy of mention is the use of $P < 0.05$ as a standard level of statistical significance. This value (essentially a 5% probability of having a false-positive result) is arbitrary. It has been established by convention and is problematic, to say the least.^{12,13} For some experiments, a P value of 0.06 may be acceptable. For others, absolute certainty that the effect is real may be necessary before one would consider changing one's practice based on the researchers' conclusions; in that case, a P value of 0.01 would be necessary. In general, it is always better for a researcher to include the absolute P values in an article and thereby let the reader judge whether the difference is significant. Furthermore, when an author states that there is no difference in a result, there should be some discussion of the power of the experiment or an inclusion of the confidence intervals (CIs). Also, if a sample size is very small, a real difference may not be detected.

Many statisticians advocate the use of CIs,¹⁴ which offer much more information than P values because CIs do not require one to assume a null hypothesis, even temporarily.^{15,16} The CI offers a range of values, with some predetermined level of confidence, that is believed to surround the true value.^{17,18} Hence, unlike the P value, which tends to segregate results as either "significant" or "not significant," the CI provides an estimate of how great the difference between two groups actually may be.^{17,18} It provides the range of true differences that are consistent with the observed data.

It is also important to be able to recognize certain characteristics of the data in an experiment. Data can be

classified as either discrete or continuous. Discrete data (or categorical data) are data that fall into separate categories. Examples of discrete data include results that are classified as poor, fair, good, and excellent, or patient satisfaction recorded as yes or no. In contrast, continuous data fall along a spectrum. Examples include height, weight, or numeric data from many of the validated outcome-assessment tools used in outcomes research. Continuous data occasionally fall on a normal distribution, which produces a bell-shaped curve that is symmetric about the mean; here the mean and standard deviation are appropriate descriptive statistics. More often, continuous variables have non-normal distributions. When the distribution of a variable is unknown or is asymmetric, descriptive statistics such as quantiles are appropriate (in fact, quantiles are appropriate for any continuous variable). Quantiles are always descriptive of continuous variables regardless of the underlying distribution, and a good three-number summary is the lower quartile (25th percentile), median, and upper quartile (75th percentile), where the median is representative of a "typical" subject.

It is important to ensure that the correct statistical test is used according to the nature of the data. In general terms, a discrete response variable is analyzed using statistical tests, such as chi-square tests, based on frequency of occurrences.¹⁹ Variables that can only be considered ordered and whose values may not be considered interval scaled (eg, pain sensitivity: none, mild, moderate, severe) should be analyzed with nonparametric rank methods that do not assume anything about the spacing of the values in the scale. Variables that can be considered interval scaled (eg, moving from a value of 2 to 3 is similar to moving from 9 to 10 in some respect) may be analyzed by either nonparametric or parametric methods. Parametric tests assume normal-

ity and typically are used when the raw data arise from a symmetric distribution; however, nonparametric tests have excellent power and can be used for normally distributed data, as well.²⁰ When the data are not normally distributed, nonparametric tests generally have greater power and are preferred over the normality-assuming tests. In addition, nonparametric tests are little affected by extreme values in the data.

For statistical analysis, raw data must be transformed (eg, the mean and standard deviation must be determined). Nonparametric tests have the advantage of yielding the same *P* value no matter how the variable is transformed, as long as the transformation preserves the rank ordering of values. Therefore, many statisticians recommend the use of nonparametric methods when a *P* value is desired and when complexities such as covariate (confounder) adjustment are not required. Regardless, the reviewer should assess the correctness of statistical tests that are used to assess the evidence of pivotal comparisons in the paper. This assessment should include (1) whether the right test was selected for the type of response variable, (2) whether paired-versus-unpaired methods were properly selected, (3) whether distributional and equal variability assumptions were strongly justified if parametric methods were used, and (4) whether proper adjustment was done for both confounding (in a non-randomized study) and significant patient heterogeneity in outcome tendencies (in any study).

Generally, the use of a statistical test that is not designed for the data (eg, using Student's *t* test when the data do not follow a normal distribution) is more conservative and will result in *P* values that are higher than would be obtained if the appropriate statistical test were used. As a result, when a poorly chosen test yields a significant result, it is more than likely that the significance is trustworthy

as long as the data were not used to select the test or to categorize continuous variables. It is not proper to attempt to use more than one test in an effort to obtain significance. Again, this offers another advantage for the use of CIs over *P* values.

Finally, it is important to note that statistical significance does not always imply clinical significance. Indeed, essentially any observed difference can be shown to be statistically significant if the sample size is sufficiently large. A study comparing two different anterior cruciate ligament grafts, for example, may show a difference in KT1000 (MEDmetric, San Diego, CA) arthrometry laxity at a *P* = 0.001 level. However, if that difference is 1 mm, we must ask whether it is meaningful to the patient.

An Approach to Reading and Interpreting the Literature

These evidence-based medicine principles may be used in a systematic way to review medical information, presentations at meetings, and manuscripts. The clinician thus may determine the likely validity of information and decide whether it should be used to change his or her practice of orthopaedics. The method described below has been modified from Tricia Greenhalgh's introductory evidence-based medicine text, *How to Read a Paper: An Evidence-based Medicine Approach*,² and from Lang and Secic's *How to Report Statistics in Medicine*.²¹ This method is not the only approach one can take; however, the authors have found it to be helpful in reviewing the literature and making decisions about how to practice orthopaedics.

Appendix 1 (available at <http://www.jaaos.org/cgi/content/full/13/4/220/DC1>) is a worksheet developed by the authors to extract evidence-based medicine information from a manuscript. The first step is to record

the citation, including the title, authors, and journal information. Next, the introduction (or, less commonly, the early part of the methods section) must be reviewed to clearly identify the research question and/or hypothesis. It is important to recognize and record the research question or hypothesis because, as readers of a paper (and reviewers of data), we are asked to judge whether the authors have designed and implemented the research appropriately, as well as whether their data address the stated research question or hypothesis. Case reports and expert opinion papers may not have a clear research question or hypothesis; nevertheless, the purpose of the manuscript frequently is stated. Any secondary questions or hypotheses also should be recorded.

The methods section of the manuscript typically is reviewed next. Here the reviewer should identify the study type. Is this study designed to test a new treatment, evaluate a new screening or diagnostic test, determine prognosis, or demonstrate an association between a potential risk factor and a disease or condition? Once the type of study is identified, the reviewer should identify the study design (which is usually, but not always, clearly stated in the methods section). The reviewer also should determine whether the appropriate study design was used for the type of study initiated (Table 2). Evidence-based medicine requires us to ask several other questions when reviewing the methods section, such as the following. Is there a control group in a treatment study? Are there other factors that the authors did not control for when comparing two groups? Is the study prospective or retrospective? Did the authors use an established gold standard for a diagnostic study?

Finally, it is important to look for potential sources of bias and to determine whether the authors have made an effort to prevent bias from entering their work. When a poten-

tial source of bias is identified, it should be listed on the worksheet. One method that helps reduce bias is the use of independent examiners in determining objective outcome data; therefore, when independent examiners are used in a study, this fact should be recorded on the worksheet. When a validated outcome measure is used, this too is recorded.

Next on the worksheet, the population of each patient group in the study is recorded. The number of patients within each group that received treatment, and the number of patients available for follow-up assessment, are recorded. The number of patients also may be recorded according to the reported results in the outcomes section for the outcomes of interest. These data are helpful in determining whether the study had adequate follow-up. These data also provide information to estimate the power of the experiment. Important questions to ask and record at this point include the following: Did the populations under comparison have similar patient characteristics? What is the length of the follow-up? Were there

any differences in the posttreatment care?

The next step is an evaluation of the statistical approach. Here some knowledge of statistics can be of great help. On the worksheet, a reference table is provided to help in determining whether the statistical test was appropriate. In addition, a detailed author checklist of statistical errors to avoid is available online from the Department of Biostatistics at Vanderbilt University School of Medicine.²² The reviewer with a limited knowledge of statistical methods should consider obtaining assistance in reviewing the statistical analysis.

The results section of the manuscript is reviewed next. For each comparison the authors set out to make, the reviewer should record the result, the point estimate of the difference between the groups, the *P* value, and, if no statistical significance was found, the power of the experiment (or a comment on whether the confidence limits were sufficiently narrow to allow any conclusions). Finally, the reviewer must decide whether the differences detected are clinically significant.

With all of this information recorded, the reviewer is ready to assign a level of evidence, using the evidence-based medicine hierarchy of evidence (Table 3). Then the reviewer or clinician may determine whether this research effort is both sufficiently valid and generalizable to change the way one practices orthopaedics.

Summary

Despite numerous conflicting opinions in the field of orthopaedics, clinicians are united in their efforts to provide the best and most up-to-date care. Using the evidence-based medicine approach, we can review the vast amount of information that we read and study, determine how valuable it is, and make rational decisions based on sound scientific principles to change our practices. This approach will help the field of orthopaedics evolve efficiently, will help direct research efforts toward producing more useful information, and will help ensure that patients receive care based on sound scientific principles.

References

1. Sackett DL, Rosenberg WMC, Gray JAM, Haynes RB, Richardson WS: Evidence-based medicine: What it is and what it isn't. *BMJ* 1996;312:71-72.
2. Greenhalgh T: *How to Read a Paper: The Basics of Evidence Based Medicine*, ed 2. London, United Kingdom: BMJ Books, 2001.
3. Bernstein J: Evidence-based medicine. *J Am Acad Orthop Surg* 2004;12:80-88.
4. Dunn WR, Lyman S, Marx R, ISAKOS Scientific Committee: Research methodology. *Arthroscopy* 2003;19:870-873.
5. Kuhn JE, Greenfield ML, Wojtys EM: A statistics primer: Types of studies in the medical literature. *Am J Sports Med* 1997;25:272-274.
6. Kuhn JE, Greenfield ML, Wojtys EM: A statistics primer: Prevalence, incidence, relative risks, and odds ratios. Some epidemiologic concepts in the sports medicine literature. *Am J Sports Med* 1997; 25:414-416.
7. Morrison DS, Bigliani LU: The clinical significance of variations in the acromial morphology. *Orthop Trans* 1987;11:234.
8. Guyatt GH, Sackett DL, Sinclair JC, Hayward R, Cook DJ, Cook RJ: Users' guides to the medical literature: IX. A method for grading health care recommendations: Evidence-Based Medicine Working Group. *JAMA* 1995;274:1800-1804.
9. Greenfield ML, Kuhn JE, Wojtys EM: A statistics primer: P Values. Probability and clinical significance. *Am J Sports Med* 1996;24:863-865.
10. Greenfield ML, Kuhn JE, Wojtys EM: A statistics primer: Power analysis and sample size determination. *Am J Sports Med* 1997;25:138-140.
11. Kuhn JE, Greenfield ML, Wojtys EM: A statistics primer: Hypothesis testing. *Am J Sports Med* 1996;24:702-703.
12. Cohen J: The Earth is round ($p < .05$). *Am Psychol* 1994;49:997-1003.
13. Goodman SN: *p* values, hypothesis tests, and likelihood: Implications for epidemiology of a neglected historical debate. *Am J Epidemiol* 1993;137:485-496.
14. Greenfield ML, Kuhn JE, Wojtys EM: A statistics primer: Confidence intervals. *Am J Sports Med* 1998;26:145-149.
15. Dorey F, Nasser S, Amstutz H: The need for confidence intervals in the presentation of orthopaedic data. *J Bone Joint Surg Am* 1993;75:1844-1852.
16. Szabo RM: Principles of epidemiology for the orthopaedic surgeon. *J Bone Joint Surg Am* 1998;80:111-120.
17. Borenstein M: Planning for precision in survival studies. *J Clin Epidemiol* 1994; 47:1277-1285.
18. Parker RA, Berman NG: Sample size: More than calculations. *Am Stat* 2003; 57:166-170.
19. Hollander M, Wolfe DA: *Nonparametric Statistical Methods*, ed 2. Hoboken, NJ: Wiley-VCH, 1999.
20. Greenfield ML, Wojtys EM, Kuhn JE: A

- statistics primer: Tests for continuous data. *Am J Sports Med* 1997;25:882-884.
21. Lang TA, Secic M: *How to Report Statistics in Medicine: Annotated Guidelines For Authors, Editors, and Reviewers*. Philadelphia, PA: American College of Physicians, 1997.
 22. <http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/ManuscriptChecklist>.