



Special Topics

Orthoepidemiology 101: The Basics of the Art of Outcomes Research in Orthopaedic Surgery

Keith D. Baldwin MD, MSPT, MPH; Rachel L. Slotcavage*, MD;

G. Russell Huffman MD, MPH

*The research fellowship of Dr. Slotcavage was funded by Stryker Orthopaedics.

Introduction

Epidemiology is defined as “the branch of medicine that deals with the study of the causes, distribution, and control of disease in populations.”¹ The outcomes research branch of epidemiology is especially important in orthopaedic surgery, as it pertains to the outcomes of treatment of traumatic and atraumatic conditions for which the natural history is often poorly understood.² Performing outcomes research is as much an art as it is a science. In spite of this, there is an orderly fashion by which outcomes research should be planned, performed, and disseminated. This article describes such an approach.

Beginning Phase

The very beginning of any project in clinical or outcomes research involves outlining some basic principles.³ First, what exactly is the question the researcher wishes to address? Second, what data is the most relevant to collect to answer that question? Third, how can a database be structured to make data analysis easy and fluid, and what analyses are best suited for the type of data the researcher has? Lastly, what is the best method by which to disseminate the results?

Address for Correspondence:

Keith D. Baldwin, MD, MPH, MSPT;
Hospital of the University of Pennsylvania, Department of Orthopaedic Surgery
3400 Spruce Street, 2 Silverstein Building
Philadelphia, PA 19104
Phone: 1-215-662-3340
Fax: 1-215-349-5890
Email: keith.baldwin@uphs.upenn.edu

Keith Baldwin is a Resident in the Department of Orthopaedic Surgery at the University of Pennsylvania

Rachel Slotcavage is a Research Fellow in the Department of Orthopaedic Surgery at the University of Pennsylvania

G. Russell Huffman is an Assistant Professor in the Department of Orthopaedic Surgery at the University of Pennsylvania

Developing a question

Often forgotten, the most important step to beginning a research project is the identification of a research question.⁴ The research question must be interesting, pertinent, and answerable with data the researcher can access, and usually requires a thorough review of the related literature. Without *a priori* knowledge of the question a researcher wishes to answer, it is difficult to collect appropriate data and, subsequently, perform a data analysis that produces meaningful information. At worst, forming a database without first determining a feasible research question will infuse insurmountable bias (systematic error resulting from flawed study design or conduct) into the study as a result of data mining. In addition, even if a meaningful research question is developed later on, further data parameters not previously collected may be necessary to fully answer the question, resulting in further expenditure of time and effort.

Study Design

After a research question has been developed, a researcher must consider which study design is most appropriate to answer the question.^{5, 6} Should the study be prospective or retrospective? Prospective studies are often appropriate in comparing interventions where there is a reasonable degree of uncertainty as to which treatment option is best. This study design allows for the largest degree of control over outside variables. Subjects can be randomized to the treatments in various ways in order to account for confounding (the chance that some factor other than the intervention is responsible for the outcome). The downside of prospective studies is that they are cost-, time-, and labor-intensive, making them frequently difficult for a resident or busy practitioner

to accomplish. Retrospective studies are often inexpensive, less labor-intensive, and provide results in a shorter time period than a prospective study. These benefits come at the cost of introducing the possible effects of confounding or bias. In many cases, these problems can be partially corrected with appropriate statistical analysis and carefully selected historical controls.⁷ Though bias and confounding are frequently cited as limitations in retrospective studies, they can also exist in poorly controlled or randomized prospective studies.

The classic example of a prospective study is the randomized controlled trial (RCT), which involves comparing multiple interventions or the utility of an intervention vs. placebo. These are considered the gold standard of studies, which, if well designed and controlled, provide the most valid and “true” results. There are, however, other prospective study designs which are easier to implement. A temporal comparison or “pre-post” study can examine the effect of a treatment(s) by examining the same population before and after an intervention, with each subject acting as their own control. A prospective cohort study involves the identification of subjects’ previous exposures by the researcher, who then follows the subjects forward in time to see who develops the outcome of interest. There are several types of retrospective studies as well. In a case-control study, cases are chosen retrospectively by outcome, and appropriate controls (matched or not) are chosen who lack the outcome of interest. Both groups are then examined for previous exposures. This type of study is often appropriate when the outcome of interest is rare, since it ensures an adequate number of cases. A retrospective cohort study is a design in which subjects are chosen for the presence or absence of a certain exposure or treatment (e.g. patients treated surgically and non-surgically for the same condition), and then examined for an outcome which has already occurred. This type of study is appropriate when the exposure or treatment of interest is not common. A case series is considered a lower Level of Evidence, but is appropriate to report the safety of a surgical technique or describe outcomes of a treatment or injury for which controls are not available. This type of study is limited because there is no evidence that the outcome, whatever it may be, would not have occurred in the absence of the treatment or injury.

The question of whether to use a study design similar to that of previous studies of the intervention or exposure or differently is a matter of what question the study is meant to answer. If the study is designed to contribute to or refute a growing body of literature, then the best design is one similar to studies performed in the past so that they can be directly

compared in a meta-analytic fashion. On the other hand, if past studies utilized a suboptimal study design or were flawed in data analysis, a researcher may wish to perform a study with a more appropriate design at the expense of being able to directly compare to previous studies

Outcome and Data Selection

The next task is to determine what type of outcome is the most appropriate to collect for a given exposure or treatment of interest.⁸ The most appropriate outcomes are patient-centered, and should be considered important to both patient and surgeon. Outcomes instruments, such as the SF-36 and TESS⁹ surveys, are appropriate if they are valid (meaning they adequately measure what they are intended to measure) and the measures they contain fulfill the aforementioned criteria.^{10, 11} Otherwise, other simpler variables such as return to work, return to sport, etc., can be used as outcomes, and are universally clinically significant.

Researchers should do their best to avoid surrogate endpoints or variables.¹² Surrogate variables are outcome measures thought to be directly linked to the outcome of interest, and are used when the actual outcome of interest is rare or takes many years to occur. The classic surrogate variable in orthopaedic research is deep vein thrombosis (DVT) as a surrogate for pulmonary embolus (PE) in researching prophylaxis for postoperative thromboembolic events. The event we actually care about (PE) is too rare to research as an outcome, so we research something thought to be directly related to the development of that outcome (DVT). However, there is currently no way to determine which patients with DVT will go on to develop PE, and many will not (DVT is necessary, but not sufficient, for PE). Therefore, despite its acceptability in clinical practice, DVT is questionable as a surrogate variable for PE. This type of issue occurs throughout orthopaedic research.

Additionally, it is important to give some thought to other data which must be collected outside the primary outcome in order to account for confounding.¹³ Possible confounders are obtained from a detailed review of the literature based on what has been shown to be clinically important in the past, or can also be based upon anecdotal clinical evidence (“In my practice, patients with characteristic A always seem to do worse.”). These variables must also be included in your statistical analysis in order to determine their true effect.

Sample Size and Power

Sample size and power analyses should be based upon the primary outcome of interest.¹⁴ For example, if the primary outcome of interest is a continuous variable (see discussion on variable types in *Database Development*, below), and the t-test will be used for analysis, then a power analysis based on the t-test should be used when planning the study. Some common sample sizes can be calculated via web based calculators.¹⁵ Many of these calculators require that you select a control for type I error or α , a control for type II error or β , and a delta value. Type I error is the chance of finding a difference where none exists. The most often tolerated is a type I error rate of 0.05, and is the same as saying that one is 95% certain that the outcome resulted from the variable of interest and was not due to chance alone. Type II error rate is the chance of not finding a difference where one actually exists. One minus the type II

error rate is referred to as power and 80% power is considered tolerable by convention. An underpowered study that finds no difference is less meaningful than an adequately-powered study that finds no difference, because in an underpowered study there is a strong likelihood that the negative result occurred by chance. Delta is the change the researcher wishes to detect, and should be considered clinically significant. It is considered less often than type I and type II error, but is of similar if not greater importance. This number is most often derived from past studies or taken from clinical experience, and is vitally important to the power calculation. The larger the delta is (the larger the change expected), the smaller the sample size would need to be. Delta must also be individualized; for example, the delta required in a power calculation for a continuous outcome variable is a continuous number, whereas the delta for a binary variable is a percentage.

Beginning steps to starting a research study:

1. Develop a question
2. Decide on an appropriate study design to answer the question
3. Decide on an outcome of interest and study variables
4. Perform a power analysis to determine how many subjects are needed

Figure 1: Steps to beginning an outcomes research project

Data Collection and Analysis

There are many different ways to collect data. The “best” way to collect data will often depend on the study question, study design, the resources available to the investigators, and the logistics of collection. Database mining, phone interviews, patient visits and chart review are all viable options. Each has its own limits and is subject to its own sets of problems. Public or private databases are a convenient source of data; however, a researcher is limited to the data fields available in the database and missing data is often not recoverable. Misclassification bias, a systematic distortion of information resulting from inaccuracy in measurement (e.g. on a survey, a patient may report a superficial wound infection after arthroplasty as an “infected total hip”), may also be present and difficult to account for. Chart reviews are a common way of obtaining data, especially in retrospective studies. Data obtained in this way may be slightly less vulnerable to misclassification due to

the medical knowledge of its authors, but missing data can still be difficult to reconcile. Phone interviews and patient visits are far less prone to missing data, but are generally more labor-intensive and may be vulnerable to responder bias (the censoring, intentional or unintentional, of information on behalf of the subject).

Database Development

Database development is a critical phase in the project, but the overall design should be pre-determined in the initial phases based upon what data is being collected. The best method in which to record the data will depend on what form the outcome of interest takes. Binary data are data in which a characteristic either does or does not exist: dead or alive, heads or tails, and surgery or no surgery are examples of binary data. Categorical data are data in which logical categories divide the data, and can be either nominal or ordinal. Nominal data

are data which have no inherent value or order, just names. For example, Zimmer™ and DePuy™ are brands of knee replacement prostheses, but have no inherent order. Ordinal data have a rank order, but there is no scale between numbers. The classic example is the pain scale: a patient who rates their pain as a 6 does not necessarily have twice the amount of pain as someone who reported a 3. Continuous data are data that are numeric and have some scale associated with them, such as temperature, range of motion, and follow-up time. There are also special types of data such as time scales and count data that have special tests associated with them, and are beyond the scope of this introductory article.

During both study design and database development, it is important to note all significant confounders, along with the independent variable of interest, and the form they will take. For example, suppose our research question was “Does cigarette smoking increase infection risk following fracture surgery?” Infection would be our dependant variable (binary: either infection or no infection), and smoking would be our primary variable of interest, and could be binary (smoker or non smoker), categorical (never smoked, current non smoker, smokes less than a pack/day, smokes greater than a pack/day), or perhaps even continuous (number of cigarettes smoked per day). The form the variable takes should be based on what past research has shown to be significant. Previously reported confounding factors must be considered, such as patient age^{16, 17} (continuous or binary [i.e. <65 years old and >65 years old]), diabetic status¹⁶ (binary), whether the fracture was open or closed¹⁶ (binary), Gustilo classification¹⁷ (ordinal), OTA classification^{16, 18} (ordinal), the location of the fracture¹⁶ (nominal), and injury mechanism or energy¹⁷ (nominal or ordinal, respectively). Continuous data always provide more information than binary data, and should be used whenever possible.

Some special statistical considerations exist in outcomes research, but are relatively minor if adequately planned for and understood in advance. The first of these which commonly occurs is the necessity of adjustment for multiple tests. The overall type I error tolerance for most studies is 0.05, but if multiple tests are conducted (i.e. one for each confounder), then there is an increased chance to commit a type I error (the more things you look for, the more you will find just by chance). To ameliorate this problem, a correction can be made for multiple tests, with the most well known and conservative of these is the Bonferroni correction.¹⁹ This correction can be performed with the usage of online calculators.²⁰ Another consideration is the frustrating

problem of missing data.²¹ Missing data can create bias if dealt with in ways which are not appropriate. Fortunately, most commercially available software packages use listwise deletion (they remove any case in which missing data is present). This method can decrease the power of a study by not using all of the available data, but avoids the bias which can devalue your study.

Data analysis

One of the most important steps in data analysis is to determine which family of statistical tests is appropriate for your data: parametric or non parametric.²² Parametric tests are distributional, which means they assume that the population that your sample was derived from has some sort of standard distribution, typically based on the normal bell curve. As such, they are not always appropriate for binary (sigmoid) or categorical (multimodal) data. Parametric tests also assume a relatively large sample size (at least greater than 20), which make them inappropriate when rare exposures or outcomes necessitating a small sample size. Other assumptions apply to specific tests, but that is beyond the scope of this paper. If the data, the sample, or the population do not fit these criteria, a non-parametric or non-distributional test is appropriate. The implicit assumption of non-parametric tests is that the population which you are using the test to make inferences about is similar in distribution to your sample, regardless of what that distribution may be. The appropriate test for a given situation is determined by the number of groups to be compared, what type of data exists in each group, and whether or not a parametric or non parametric test is appropriate based on distribution and sample size. A summary of tests in various situations is outlined in Table 1. This table by no means captures every possible situation, but summarizes many of the tests which are useful in outcomes research. Multiple variables can be tested by using various statistical procedures. Much like those available for sample size and power calculations, various web based calculators exist to help calculate t-tests^{23, 24}, Chi squared tests, and Fisher's exact tests²⁵. Most if not all of the tests outlined on Table 1 can be calculated with commercially available software such as SPSS (SPSS Inc.; Chicago, IL.), SAS (SAS Institute; Cary, NC) or STATA (Statacorp; College Station, TX). In addition, many of these tests can be performed with Microsoft Excel (Microsoft Corp; Redmond, WA). Most simple tests can be calculated and interpreted by anyone with a cursory understanding of statistics. However, multiple regression analyses should be performed or reviewed by someone with advanced

Number of groups	Type of independent variable	Parametric test	Non parametric test	Multiple variables
2 groups (related)	Continuous	Paired t-test	Wilcoxin, sign test	Logistic regression *
2 groups (independent)	Continuous	Independent samples t-test	Mann-Whitney U	Logistic regression *
3 or more groups	Continuous	One way ANOVA	Kruskal Wallis,	Linear regression †
Continuous variable	Continuous	Correlation (Pearson)	Correlation (spearman)	Linear regression
2 or more	Nominal or binary or categorical	Chi squared test	Fisher’s exact test	Logistic regression*
2 or more	Time series	Parametric life table analysis	Kaplan Meier log rank	Cox regression

Table 1: Appropriate usage of various statistical tests.

*dichotomous dependant † if the dependant is continuous, multinomial regression if the dependant is categorical with greater than two categories.

training or understanding in statistics, and many of these procedures require post-hoc analyses to assure that the test procedures are valid. One specific output worth special mention in outcomes research is the Odds Ratio (OR). Odds ratios are often used to compare two groups, one that was exposed to a certain treatment, injury, or exposure, and one that was not, and to examine them both for the presence of a certain outcome. The result is expressed in terms of the odds of observing the outcome in the exposed

group compared to the unexposed group. The equation that describes the odds ratio is odds of an event in the exposed over odds of an event in the unexposed, based on the standard table shown in Figure 2. Incidentally, odds ratios which are adjusted for multiple confounding variables can be generated using multiple logistic regression, with the assistance of someone with statistical experience.

	Disease	No Disease
Exposed	A	B
Unexposed	C	D

Figure 2- Contingency table for generating odds ratios.

Presenting the Data

Presentation of the data is at least as important as any other phase of the project. A description of your statistical analysis detailed enough to allow another group to perform a confirmatory study should be included in your Methods and Materials section. A well presented graph or table placed in your Results is invaluable in getting the point of the study across. Pictures and visual aids presented in a paper enhance its interest to the orthopaedic community at large.

The decision of where to try to publish the paper is sometimes difficult, but a preliminary target journal should be identified during your initial study planning. This does not prevent having to submit to

more than one journal prior to eventual publication. Negative studies are known to be more difficult to have accepted for publication, but often contain invaluable information which should be disseminated. Statistically insignificant studies are faced with this same bias, but it is important for authors and reviewers alike to realize that statistical significance does not equal clinical significance. Few would make the distinction between a 95% and 94% certainty that your results were not due to chance alone ($p = 0.05$ and $p = 0.06$, respectively) if you are presenting them with a successful treatment with minimal morbidity for their patient population. Consideration must also be given to whether or not the article would be interesting to a more general orthopaedic audience, or if it is more specific to a

subspecialty and would thus be more appropriate for a journal within that subspecialty. Sometimes the article may be appropriate to an audience which is broader than the orthopaedic community, and an outside journal should be considered.

Conclusions

Embarking on a research project can be a long and arduous task. However, careful planning at the study outset can ensure the fruits of this task are a long lasting contribution to the orthopaedic literature and hence clinical practice. Thorough planning and meticulous execution of the research plan will minimize the chances the contribution will be tainted by confounding and bias. This is particularly important in retrospective research in which special care must be taken to limit bias and control for confounding with special statistical techniques.

References

1. Farlex. The Free Dictionary. <http://www.thefreedictionary.com/epidemiology>. Accessed October 14, 2008.
2. Novak EJ, Vail TP, Bozic KJ. Advances in orthopaedic outcomes research. *J Surg Orthop Adv*. Fall 2008;17(3):200-203.
3. Hulley SB, Cummings SR, Browner WS, Grady D, Newman TB. *Designing Clinical Research*. Third ed. Philadelphia: Lippincott Williams and Wilkins; 2007.
4. Lipowski EE. Developing great research questions. *Am J Health Syst Pharm*. 2008;65(17):1667-1670.
5. Lim HJ, Hoffmann RG. Study design: the basics. *Methods Mol Biol*. 2007;404:1-17.
6. Audigé L, Hanson B, Kopjar B. Issues in the planning and conduct of non-randomised studies. *Injury*. April 2006;37(4):340-348.
7. Dunn WR, Lyman S, Marx R, ISAKOS Scientific Committee. Research methodology. *Arthroscopy*. 2003;19(8):870-873.
8. Suk M, Norvell DC, Hanson B, Dettori JR, Helfet D. Evidence-based orthopaedic surgery: what is evidence without the outcomes? *J Am Acad Orthop Surg*. 2008;16(3):123-129.
9. Davis AM, Wright JG, Williams JI, Bombardier C, Griffin A, Bell RS. Development of a measure of physical function for patients with bone and soft tissue sarcoma. *Qual Life Res*. 1996;5(5):508-516.
10. Haywood KL. Patient-reported outcome II: selecting appropriate measures for musculoskeletal care. *Musculoskeletal Care*. 2007;5(2):72-90.
11. Haywood KL. Patient-reported outcome I: measuring what matters in musculoskeletal care. *Musculoskeletal Care*. 2006;4(4):187-203.
12. Boissel JP, J.P. C, Moleur P, Haugh M. Surrogate endpoints: a basis for a rational approach. *Eur J Clin Pharmacol*. 1992;43(3):235-244.
13. Hammal DM, Bell CL. Confounding and bias in epidemiological investigations. *Pediatr Hematol Oncol*. Sep 2002;19(6):375-381.
14. Biau DJ, Kerneis S, Porcher R. Statistics in Brief: The Importance of Sample Size in the Planning and Interpretation of Medical Research. *Clin Orthop Relat Res*. 2008;466:2282-2288.
15. Lenth RV. Java Applets for Power and Sample Size. Computer software. Available at: <http://www.stat.uiowa.edu/~rlenth/Power>. Accessed October 14, 2008.
16. Rightmire E, Zurakowski D, Vrahas M. Acute Infections After Fracture Repair: Management With Hardware in Place. *Clin Orthop Relat Res*. 2008;466:466-472.
17. Bowen TR, Widmaier JC. Host Classification Predicts Infection after Open Fracture. *Clin Orthop Relat Res*. 2005;433:205-211.
18. Castillo RC, Bosse MJ, MacKenzie EJ, Patterson BM, LEAP Study Group. Impact of smoking on fracture healing and risk of complications in limb-threatening open tibia fractures. *J Orthop Trauma*. Mar 2005;19(3):151-157.
19. Sankoh AJ, Huque MF, Dubey SD. Some comments on frequently used multiple endpoint adjustments methods in clinical trials. *Stat Med*. 1997;16:2529-2542.
20. Bonferroni adjustment online. <http://www.quantitativeskills.com/sisa/calculations/bonfer.htm>. Accessed October 14, 2008.
21. Graham JW. Missing Data Analysis: Making It Work in the Real World. *Annu Rev Psychol*. July 24 2008 [Epub ahead of print].
22. Dawson B, Trapp RG. *Basic and Clinical Biostatistics*. Fourth ed. New York: McGraw-Hill; 2004.
23. GraphPad QuickCalcs: t-test calculator. <http://www.graphpad.com/quickcalcs/ttest1.cfm>. Accessed October 14, 2008.
24. Independent Groups T-Test for Means Calculator. <http://www.dimensionresearch.com/resources/calculators/ttest.html>. Accessed October 14, 2008.
25. GraphPad QuickCalcs: analyze a 2x2 contingency table. <http://www.graphpad.com/quickcalcs/contingency1.cfm>.