ISAKOS Scientific Committee Research Methods Handbook

A Practical Guide to Research: Design, Execution, and Publication

Editors: Jón Karlsson, M.D., Ph.D., Robert G. Marx, M.D., M.Sc., F.R.C.S.C., Norimasa Nakamura, M.D., Ph.D., and Mohit Bhandari, M.D., Ph.D., F.R.C.S.C.

FOREWORD

Why Is a Research Methods Handbook Needed?

Why is this work needed, and who would benefit from it? First of all, we must realize that this work is on a high but at the same time moderate level. The aim is to put together a Research Methods Handbook that can be of practical help to those writing manuscripts for submission to *Arthroscopy* and similar journals. We are referring to people working full time, taking care of patients, with busy outpatient clinics and fully booked surgical schedules. These are persons who do not devote the majority of their time to research. And in most cases they do not have any major training in scientific research methods. Since sound research methods are the backbone of a good study, the methods must be solid to ensure that the results are valid. If the methods are not good from the beginning, the outcome will not be good either, and the manuscript will not be published despite the investigator's best effort.

The purpose of this Research Methods Handbook is to provide basic information about common research techniques, how to conduct a good study, how to write a manuscript and, we hope, how to get it published.

The work is divided into several sections, starting with an overview on evidence-based medicine; much-needed information for all clinicians. The second section is concerned with study methods, with special focus on study designs. Important scientific methods, like CONSORT and STROBE, are explained in greater detail. The third section is on biostatistics. This section is very practical, written with the clinician in mind. Common statistical methods are explained and the aim is to stay practical and pragmatic. We are still clinicians and not statisticians. And the idea is to help clinicians who are conducting a study and not to make them statisticians. The last section is on manuscript writing. Pearls and pitfalls are discussed and tips are given. We dare say that if you follow these simple guidelines, you will have a much greater chance of getting your manuscript published.

A few words of thanks. First and foremost we thank Michele Johnson, ISAKOS Executive Director, who helped out with all practical details and negotiated all necessary contracts. At *Arthroscopy*, Managing Editor Hank Hackett and Jason Miller from Elsevier made things happen. Special thanks to Hank for his professional editing work on all chapters, keeping track of the time frame, and all other practical details.

This work is an ISAKOS project, done on behalf of the ISAKOS Scientific Committee, and we would like to thank all Committee members, many of them co-authors, for their part in getting this done. Special thanks to Mario Ferretti, Stephan Lyman, Rob LaPrade, Bruce Levy, Nick Mohtadi, Kevin Shea, Michael Soudry, and Stefano Zaffagnini. We also extend our thanks to all other co-authors, with special thanks to Sabine Goldhahn. Mohit

The authors report no conflict of interest.

Address correspondence to Jón Karlsson, M.D., Ph.D., Department of Orthopaedics, Sahlgrenska University Hospital/Mölndal, SE-431 80 Mölndal, Sweden. E-mail: jon.karlsson@telia.com.

0749-8063/1194/$36.00
doi:10.1016/j.arthro.2011.02.001

Bhandari, one of the greatest clinician researchers we have ever met, deserves special thanks; without his work, this project would never have been possible.

Finally, Gary Poehling and James Lubowitz, Editor-in-Chief and Assistant Editor-in-Chief of *Arthroscopy*, supported the project from the start and shared their knowledge and vast experience in the section on manuscript writing. Thank you both. We hope that the readers of *Arthroscopy* as well as other journals will benefit from this work.

JÓN KARLSSON *Chair*
ROBERT G. MARX *Co-chair*
NORIMASA NAKAMURA *Co-chair*
ISAKOS Scientific Committee

FREDDIE H. FU
President of ISAKOS

SECTION 2

What Is This Evidence-Based Medicine and Why Bother?

The *British Medical Journal* recently surveyed the global medical community to determine the greatest medical breakthroughs since its first publication in 1840.[1] It was an incredible period of innovation and change, when antibiotics were discovered, entire joints were replaced with anything from ivory to stainless steel, internal imaging was developed allowing surgeons to see inside the body noninvasively, and vaccines were developed and implemented on a global scale. Evidence-based medicine (EBM) was noted as 1 of the top 15 medical breakthroughs in the last 160 years.

BIAS DETECTIVES

Many have compared the use of evidence in medicine to the use of evidence in the legal setting.[2] Let us consider the classic character Detective Sherlock Holmes and a legal example to set the stage.

> *You are a detective called to a robbery of a local corner store. As you are on the way to the site of the crime, you consider the last robbery you investigated at a store on the other side of town. Reminding yourself of how the last crime was conducted, you proceed to develop a theory as to how the current robbers entered the store, their path throughout the store, what they stole, and how they escaped. Yes, that must be how it happened; as you arrive at the scene, you have already pieced together the majority of the case. But what about this footprint? Does that change your hypothesis as to what went on . . . ?*

Now let's consider instead that you are this same detective but have since watched a Sherlock Holmes mystery video and have taken some of his words to heart.

> *You are en route to the site of this same robbery. While driving there, you try to clear your mind of the robbery you investigated last week. You want to approach this new case with no preconceived ideas or theories. As Sherlock said, "Never guess. It is a shocking habit . . . ."[2] You arrive at the site of the crime and begin locating evidence: a black glove here, a footprint there, a broken window in the front of the store, and a wide-open door at the back. You attempt to collect all the evidence you can find before developing a hypothesis as to the events of the robbery. Your mind recalls a quote from the detective video the night before, ". . . don't theorize before you have all the evidence."[2] Remembering how observation was second nature to Holmes, you ensure you collect all the facts and record all that you observe even if the information does not appear immediately relevant. Now it's just a matter of sitting down with the evidence and solving the crime.*

Which one of these approaches would stand up better in court? Which one would the store owner be happiest about in terms of having justice served?

REFRAMING THE PARADIGM TO EVIDENCE-BASED MEDICINE

These examples aim to illustrate, in albeit rudimentary terms, the paradigm shift that our field has been undergoing for the past decade. Over the last several years, medical professionals and health professionals have begun using EBM in their practice: integrating best available research and their clinical expertise with the specific patient's values.

The first steps of EBM (the evidence) are very similar to steps used in detective work as shown in the second example. This section will introduce the methods with which to approach a problem and track down evidence. The medicine piece of EBM is where things change. When it comes to solving the problem with the evidence you have gathered, one could argue that medicine in fact has better tools at hand than a detec-

tive would have available. This chapter will explore these tools. Lastly, applying this solution based on the evidence you have gathered for a patient's specific scenario has no parallels to detective work; this is where our clinical expertise really comes into play.

What Is Meant by Best Evidence?

If research is going to be used as evidence put toward a hypothesis, which will in turn be applied to a clinical scenario, it should aim to be best evidence. This research has to be relevant with regard to content but also with regard to what type of patient is being considered. This can range from research in basic science to patient-centered clinical research, from efficacy and safety of therapeutic regimens to the power of certain prognostic markers. The most updated clinical research does more than simply suggest new approaches, it can in fact often invalidate older diagnostic tests and treatments and replace them with ones that are more efficacious, powerful, and accurate and safer.

What Is Meant by Clinical Expertise?

Even if research can invalidate older tests and replace them with newer tests, in terms of the best approach, nothing can replace clinical expertise. Without a clinician's ability to use his or her skills and past experiences to identify the issues and a patient's health status and diagnosis, as well as risks and benefits present in the scenario, we would be hard pressed to have a starting point at which to apply mounting evidence from meetings, symposia, and peer-reviewed journals.

What Is Meant by Patient Values?

Each patient is a unique person with his or her own expectations, concerns, priorities, and preferences. When each patient brings these unique components to the clinical encounter, it is not in vain. For clinical decisions to be relevant and able to serve this particular patient, his or her unique considerations must be integrated into the decision-making process.

EBM THROUGH THE AGES

The term evidence-based medicine, a medical practice paradigm first introduced in the early 1990s, first came to light as a component of the medical residency program at McMaster University in Hamilton, Ontario, Canada.[3] What started as the introduction of "enlightened skepticism" to a group of residents at McMaster University led to an explosion of research extending this initial idea to many specialties in medicine and across the world, including orthopaedics at an international level. The methodology of EBM has become a key component of orthopaedics with journals such as *Arthroscopy*, *The Journal of Bone and Joint Surgery*, *Indian Journal of Orthopaedics*, *Clinical Orthopaedics and Related Research*, and *Acta Orthopaedica* embracing evidence-based orthopaedics as standard vernacular in their proceedings.

The concepts we now consider associated with the paradigm of EBM may have roots in ancient historical accounts of authoritative teaching and passing on of stories in ancient times or the emergence of personal journals and the introduction of textbooks in Renaissance times.[4] In the early 1990s knowledge began to be shared more easily in textbooks, and peer-reviewed journals began making an appearance in the field with regard to clinical practice. It was in the 1970s when a shift in modern technology and essentially an explosion in the field of informatics led to the emergence of online journals and large databases.

Claridge and Fabian[4] provide a variety of specific examples of EBM emerging through the ages in their 2005 report on the history of EBM. These examples indicate a gap in knowledge and a subsequent question, an approach to finding evidence, and an answer to the clinical query based on said evidence. Some of these examples are summarized in Fig 1.

Early Evidence in Orthopaedics

During the time of these more recent developments, the orthopaedics community was in the midst of developing its own evidence in the same way. Hoppe and Bhandari[5] present an interesting example of early evidence in orthopaedics by discussing a particular report from the Proceedings of the American Orthopaedic Association in 1889[6] in their article on the history of evidence-based orthopaedics. This report, entitled "Hypertrophy of One Lower Extremity," includes a case study regarding the treatment of a 6-year-old child with a leg three-quarters of an inch longer than the other.[6] After failing to slow the growth of this leg using a rubber bandage, the surgeon suggested a shoe lift for the patient's comfort. However, after the patient had later been examined by another surgeon, who diagnosed him with congenital occlusion and dilation of the lymph channels, amputation was recommended and carried out. After publication of this case, a discussion with other specialists ensued. One surgeon described a similar leg-length discrepancy presentation in a 21-year-old woman. After consultation with a colleague who was also unsure of the

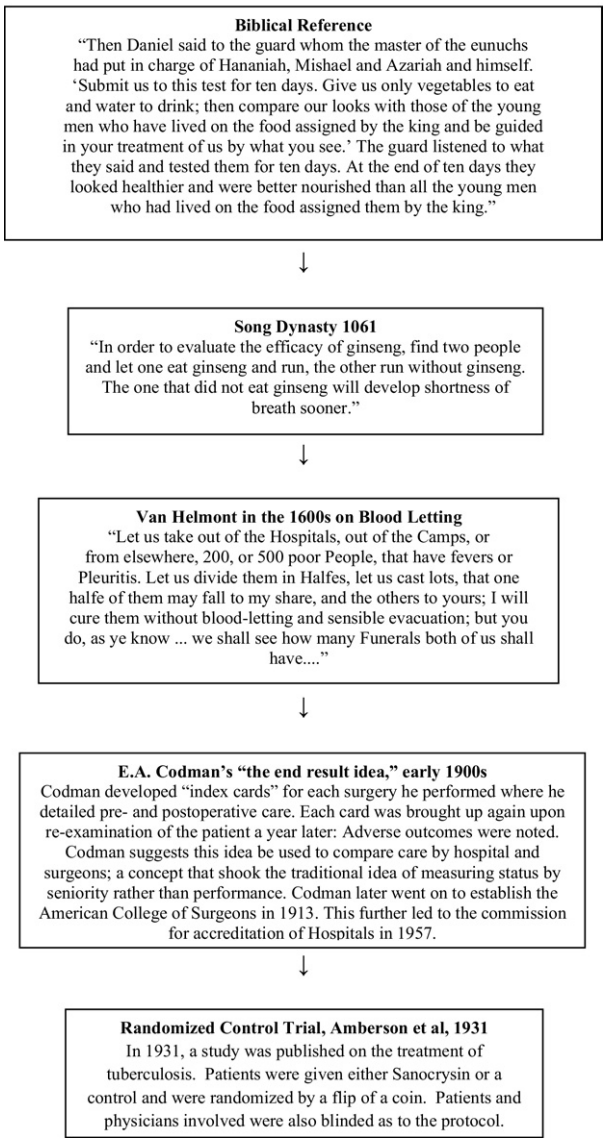

FIGURE 1. Examples of EBM through the ages adapted from Claridge and Fabian.[4]

nature of the problem, a high shoe was also given to the patient. A third case was brought up by yet another surgeon where a similar presentation was treated by stretching of the sciatic nerve.

With 3 experts offering 3 very different opinions as to how to proceed with such a presentation ranging from a shoe lift to sciatic nerve stretching to amputation, how were readers expected to know which approach to use themselves? Historically, as in many other specialties, information obtained on a case-by-case basis by experts was passed on to other doctors and learners who, knowing the expert's reputation well, would often implement a given treatment of technique into their practice with a reinforced understanding of its value.

Despite these differing expert opinions undoubtedly being a common scenario in all specialties at this time, one contributor suggested a new approach to this lack of a conscience. "Would it not be in accordance with the purposes of this association to appoint a committee to investigate this subject, taking patients . . . and treating them."[6] This is an early example of anecdotal evidence no longer being sufficient as evidence on which to base patient treatment. It was instead determined that larger-scale trials would allow these surgeons to objectively identify the superior treatment and to demonstrate the benefits of one approach versus the next.

Modern-Day EBM

From hearsay practices in ancient times to the first appearance of the randomized controlled trial (RCT) in the early 20th century and from anecdotal evidence to the development of evidence through trials in many specialties including orthopaedics, we arrive at what can be referred to as modern-day EBM.

In the early 1970s, Cochrane[7] criticized the lack of reliable evidence behind a plethora of health care interventions commonly accepted at the time. Rigorous evaluation of these interventions highlighted the need for an increase in evidence in medicine after this publication, planting the seed for EBM. David Sackett of McMaster University used the term "critical appraisal" to describe extracting evidence from systematically examined medical literature in the early 1980s.[8]

The actual term evidence-based medicine was coined by Dr. Gordon Guyatt of McMaster University in 1990. Initially a term intended for educational use by internal medicine residents in McMaster's innovative residency program, EBM gained popularity with physicians and residents in a variety of subspecialties.[9] An initial group of physicians from McMaster with a particular interest in critical appraisal grew to include specialists from a variety of institutions who joined forces to create the Evidence-Based Working Group. This group became responsible for adopting the idea of EBM and presenting it in the pivotal report announcing it as a new medical paradigm: "Evidence-Based Medicine: A New Approach to Teaching the Practice of Medicine."[10]

Emergence of EBM

There were many specific changes during this time that really set the stage for the rapid widespread rec-

ognition of EBM. There were realizations of gaps in clinical decision making preceding the coining of EBM, creating a real need for this paradigm shift. Alongside this, recent developments in technology and perspectives fostered an environment where EBM was really able to blossom in tackling these gaps.

As we approached the era of the Evidence-Based Working Group introduced earlier, it was becoming more and more evident that traditional venues for information were no longer sufficient. Several specific realizations set the stage for this spread of EBM.

Regardless of specialty, all physicians and surgeons have a continuous need for valid information on diagnosis, therapy, prevention, and prognosis for numerous patients with countless conditions and circumstances. Covell et al.[11] suggested on the basis of their research, that new information is needed 2 times for every 3 outpatients. Another study, performed in 1991, added to this suggestion stating that physicians may require new information up to 5 times per inpatient.[12]

Regardless of the increasing need, the tools and skills surgeons have typically been left with once in practice are no longer sufficient to acquire information as needed. In the past, traditional venues for finding information such as medical and surgical textbooks have been based on "expert opinion" rather than research and are in fact frequently wrong. The volume of information in these sources combined with the variability in their validity makes them an overwhelming source of information to sift through. In addition, outside of written information, traditional didactic teaching is also often ineffective when it comes to translating knowledge into clinical practice. All of this aside, for many clinicians, the main barrier to engaging in best research to find answers to clinical questions is time. With a busy clinic, operating room time, and call schedule, finding time to sit down, search for resources and assimilate information to study any given topic has often fallen outside the scope of most surgeons' typical daily schedules.

There have been various recent developments that have really allowed these previously insurmountable issues to be tackled, allowing EBM to become a day-to-day reality for full-time clinicians and surgeons. New strategies for searching for and appraising research, alongside an increase in the quality and the availability of information, have brought evidence-based practice to the forefront. In addition to the increase in amount of information, we must also acknowledge the increases in quality of research. When improvements in research are considered, a few main examples stand out. This includes an increase in recognition of the importance of clinical studies and an increased need for objective, informed consent from patients, as well as a trend for establishing globally based gold standards for best medical practice.[13] A study by de Solla Price showed that there has been an increase in the number of scientific journals by 7% per year. At this rate, the number of journals has doubled every 10 to 15 years, suggesting that by the early 21st century, we were approaching a total of 60,000 to 70,000 journals worldwide, of which 15,000 were strictly biomedical.[14]

Although this seems like an insurmountable amount of information, developments in technology have led to programs that can bring the newest valid, reliable research from a variety of sources in a concise format in a matter of seconds. The availability of systematic reviews, medical databases, the Cochrane Library, and evidence-based journals, for example, focusing on articles of immediate clinical use, has brought best research and clinical decision making closer than ever. For example, in 1997, when the National Library of Medicine announced it was offering free access to the first-line Web-based medical databases MEDLINE and PubMed, usage jumped 10-fold, to a total of 75 million searches annually.[15] Availability and accessibility of information have also increased with the advent of second-line databases such as the Cochrane Library, UpToDate, and Best Evidence along with EBM-related journals such as the *ACP Journal Club* and *Evidence-Based Medicine* (these resources will be detailed further later on). These changes, alongside the emergence of the idea of lifelong learning, explain why there has been such a sudden surge in the concept of EBM not only in theory but also in practice.

THE PRACTICE OF EBM

As discussed earlier, a doctor's clinical competence is the combination of 3 main aspects: knowledge, technical/clinical skill, and the ability to make decisions. The cumulative factor of this combination is the ability to make appropriate, systematic, and unbiased decisions to predict prognosis and interpret the results of examination and laboratory tests to overall achieve therapeutic efficacy. In 1995 Haynes and Sackett[16] summarized the key steps of practicing EBM in the opening editorial of the journal *Evidence-Based Medicine* as follows.

1. Formulate the problem and convert the information needs into answerable questions.
2. Search for and assimilate in the most efficient way possible, the best evidence with which to answer these questions. This information comes

from the clinical examination, laboratory tests, diagnostic imaging, or published literature.
3. Once collected, appraise the evidence critically for both its validity and its applicability to the current clinical question.
4. Apply the results of this search and appraisal in practice to both the clinical question and patient context.
5. Evaluate the above steps.

These steps will be outlined in further detail, illustrating how surgeons taking as little as 30 minutes of time per week for their professional development can implement EBM into their practice to answer anything from the everyday common complaint to the less common complaint to the rare presentation.[17]

EBM AT WORK

Knee pain is among the most common complaints of patients seen by both primary care physicians and orthopaedic specialists. Despite how often this type of patient presents, many clinicians still struggle with evaluating knee pain. After developing a clinical picture through discussion of the nature of the pain, mechanism of injury, patient's history, relevant physical examination findings, and preliminary diagnostic imaging, many clinicians are still unsure of how to proceed with regard to further investigation. With no clear diagnosis at this point, does this patient need a magnetic resonance imaging (MRI) scan? Will this add to the clinical picture, or will this not provide any new information? When it comes to efficiency and economy of practice and allocation of resources, being able to determine whether an MRI scan is required in this presentation is essential. Ordering an MRI scan because you "always do" or because a mentor has always suggested it is no longer sufficient evidence to warrant proceeding.

Recent research by Matzkin et al.[18] presented at the American Association of Orthopaedic Surgeons meeting in 2011 has produced an evidence-based algorithm to determine the need for an MRI scan in evaluation of knee pain. By considering duration of symptoms, presence of an effusion, laxity, joint-line tenderness, and the degree of radiographic degenerative change, this algorithm will indicate the need for an MRI scan in this situation. This algorithm is an excellent example of how evidence derived from well-conducted, valid, and reliable research is coming to the surface as we speak, influencing our standard of care in the most common presentations.

Asking a Well-Built Research Question

As mentioned, formulating, building, and focusing a clinical question comprise the first step in an approach to using EBM in practice. Every time we see a patient, for the first or fifth time, there is a need for new information about some component of our approach: the presentation, diagnosis, prognosis, or management. These gaps in knowledge combined with our limited time to devote to research necessitate a focus on efficiency. Our gaps in knowledge can sometimes seem rather large, so with this in mind, alongside our limited time to devote to this, we must be as efficient as possible in our search. The first key factor in keeping this step efficient is to become skilled at formulating answerable clinical questions.

Questions commonly arise regarding anything from clinical findings, differential diagnoses, manifestations, harm, and etiology to therapy, prevention, diagnostic tests, and prognosis. Examples of such common questions are shown in Table 1.

Components of a Good Question

1. The patient context, problem, or population in question
2. The potential intervention, exposure, or maneuver
3. The approach/option to which this intervention is compared
4. Clinical outcome of combining the above 2 factors considered in a specific timeline

These 4 components, identified with the acronym PICO, are detailed below.

Patient Characteristics: To set a good context for any question, clinicians must first identify and consider

TABLE 1. *Common Questions*

Harm/etiology: Questions of identifying and understanding the cause for a condition or disease.
Prevention: Questions related to reducing the chance of a disease developing. This involves identifying and understanding modifiable risk factors associated with the condition as well as early screening techniques and standards.
Diagnostic test: Questions related to selection and interpretation of diagnostic tests and, from this, how to confirm or exclude a diagnosis. This involves consideration of a test's specificity, sensitivity, likelihood ratios, cost, risks, and so on.
Therapy: Questions related to selecting appropriate treatments, weighing the associated risk/benefits, and efforts/costs of using them.
Prognosis: Questions related to estimating the likely clinical course for a given patient over time and any complications associated with this.

the patient's characteristics. This involves demographic information such as age, sex, and race, alongside their social situation, resources, and values. In addition to this demographic information, characteristics specific to the clinical situation such as diagnoses or condition must be included. The setting (inpatient, outpatient, rural, tertiary care, and so on) must then be considered. Is this a public health issue or an individual patient issue?

Intervention: In forming a well-built clinical question, the intervention must then be included. What is it exactly that is being considered as a potential intervention? This could be a medication, a diagnostic test, or any other type of treatment.

Comparison: A treatment or test can really only be assessed relative to or in comparison with something else. One side of this comparison will be the potential intervention, and the other will be that against which it is being compared. This comparison may be another test or treatment, the current standard treatment, watch and wait, or even no treatment at all.

Outcome: Once the above are determined within the clinical question, include the outcome as well. What is the desired effect you want to achieve? Is there an effect you want to avoid? This can involve not only treatment effects but also side effects. Outcome will typically be divided into a primary outcome and surrogate outcomes (measurements that on their own hold little value for the patient but are associated with outcomes that are considered very important to patients).

Instead of asking, "Is operative treatment indicated for a fractured clavicle?" ask, "In an active adult patient with a completely displaced midshaft clavicular fracture, would primary plate fixation result in improved functional outcome when compared with nonoperative treatment at 1 year of follow-up?"

By using the PICO model to develop a specific and complete clinical question, the task of finding best evidence becomes more plausible and efficient.

Finding the Evidence in the Literature

Developing techniques for searching for evidence may seem daunting. Considering that MEDLINE adds 4,500 records to its database on a daily basis, a physician in any one field would need to read 18 articles per day, 365 days a year, to be able to keep up with this amount of research[11]: hence, daunting. This type of reading schedule is not plausible for any busy clinician or surgeon. Add to this that, in fact, only 10% of these articles are considered to be high quality and clinically relevant, and this task seems even less plausible.[11] In reality, however, by learning how to effectively approach a search for evidence, learning where to look and what techniques to use now, this job on a day-to-day basis becomes increasingly less daunting. In this section we will discuss various key concepts, tips, and approaches to develop ways to find the evidence in an efficient and effective way.

When first approaching the vast and continually growing number of scientific and medical articles available, an easy first step is to understand and identify the different types of research study designs. Descriptions of the different type of research study designs are listed in Table 2.

From here, the types of research are placed in a hierarchy based on their value. Figure 2 illustrates the pyramid, or hierarchy of evidence, that reflects the

TABLE 2. *Study Designs Defined*

Meta-analysis: A combination of all of the results in a systematic review using accepted statistical methodology.
Systematic review: On the basis of a specific clinical question, an extensive literature search is conducted identifying studies of sound methodology. These studies are then reviewed, assessed, and summarized according to the predetermined criteria related to the question at hand.
Randomized (clinical) control trial: A prospective, analytic, experimental study that uses data generated typically in the clinical environment. A group of similar individuals are divided into 2 or more groups (1 acting as a control and the other[s] receiving the treatment[s]) and the outcomes are compared at follow-up.
Prospective, blind comparison to a gold standard: To show the efficacy of a test, patients with varying degrees of an illness undergo both the test being investigated and the "gold standard" test.
Cohort study: A large population with a specific exposure or treatment is followed over time. The outcomes of this group are compared with a similar but unaffected group. These studies are observational, and they are not as reliable because the 2 groups may differ for reasons aside from the exposure.
Case-control study: Patients who have a specific outcome or condition are compared with those who do not. This is a retrospective approach used to identify possible exposures. These are often less reliable than RCTs and cohort studies because their findings are often correlational rather than causative.
Case series/report: Reports on the treatment of an individual patient are reviewed. These have no statistical validity because they use no control group for comparison. Case reports do, however, have a role for novel and rare presentations, because no large populations exist in these cases.

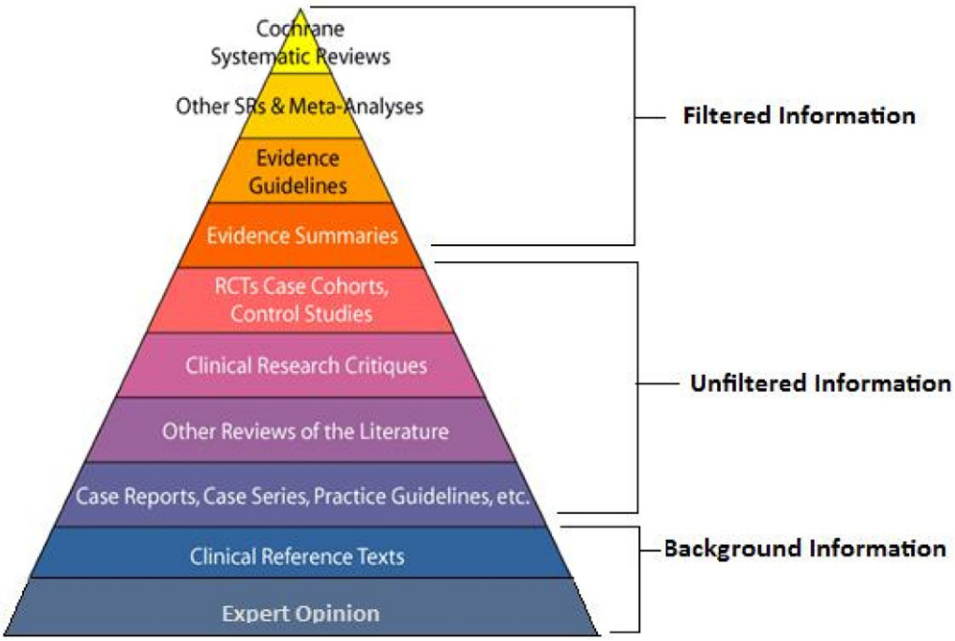

FIGURE 2. Hierarchy of evidence. This image separates the different types of research into 3 categories: background information, unfiltered information, and filtered information.

relative authority of the different types of research present in the biomedical field. It is important to note that although there are various versions of this hierarchy and there is no universally accepted version, there is still agreement on the strength of certain key types of research relative to the others. By understanding how different types of research compare to one another, those that are most useful for a busy practicing clinician with a specific question can be targeted. As you move up this pyramid, the quality of the research increases, in that it is most relevant to the clinical setting and has a reduced risk of bias compared with modes of research lower down the pyramid. In addition, research higher up the pyramid puts less onus on the searcher with regard to filtering through original data, making such research a much more efficient means of locating information.

Filtered Resources: With a clinical question related to the course of action/management of a patient, be it related to diagnosis, treatment, prognosis, and so on, filtered resources should be consulted first. Filtered resources (examples of which are shown in Table 3) consider a question posed by clinical experts and topic specialists and then provide a synthesis of evidence to come to a conclusion based on all available research. Using filtered information is much more efficient because the searching clinician does not need to individually appraise each piece of evidence. The clinician still has a responsibility to evaluate the information with regard to the specific patient and context in question. To aid with this portion, these resources also back up information with links to the relevant literature and resources. When searching in Ovid and PubMed, clinical filter options can be applied to aid in finding EBM research.

Unfiltered Resources: If an appropriate answer to the clinical question is not found in the filtered resources, the primary literature or unfiltered resources must be considered. Unfiltered resources also provide the most recent research and can be used to determine

TABLE 3. *Examples of Filtered Resources*

- Systematic reviews and meta-analyses
  - Cochrane Database of Systematic Reviews (The Cochrane Collaboration)
  - Database of Abstracts of Reviews of Effects (DARE; National Institute of Health Research)
- Critically appraised topics (evidence syntheses)
  - Clinical evidence
  - InfoPOEMs (Canadian Medical Association)
  - ACP PIER (Physician's Information and Education Resource; American College of Physicians)
  - National Guideline Clearinghouse (Agency for Healthcare Research and Quality)
- Critically appraised individual articles (article synopses)
  - Evidence Updates
  - Bandolier
  - ACP Journal Club

TABLE 4. *SORT Rating System*

Code	Definition
A	Consistent, good-quality patient-oriented evidence
B	Inconsistent or limited-quality patient-oriented evidence
C	Consensus, disease-oriented evidence, usual practice, expert opinion, or case series for studies of diagnosis, treatment, prevention, or screening

whether any new strides have been made in this area since the conclusions in the filtered resources were released. The challenge with unfiltered resources is that the onus is put on the clinician to evaluate each study to determine its validity and applicability to the query at hand. Searching for these resources efficiently and subsequently appraising what is found take more time and skill, which is why filtered information is typically considered first.

MEDLINE is considered the database of choice for the health sciences because it provides both primary and secondary literature for medicine and other allied health professionals. In these instances, RCTs, meta-analyses, and systematic reviews are considered the gold standard and should be considered first.

Ratings of Quality of Evidence: Various rating scales have been developed to help the busy clinician gauge the quality of research based on an externally applied rating before starting the critical appraisal step of practicing EBM. The Centre for Evidence-Based Medicine in Oxford provides 3 different rating scales ranging from 1 to 5, each number and occasionally added letter identify the level of evidence based on type of research design and various measures of quality, such as confidence intervals and randomization. The most updated of these detailed scales can be accessed at www.cebm.net.

Strength of Recommendation Taxonomy (SORT) (with codes A, B, and C) is a straightforward rating system,[19] shown in Table 4.

Grading of Recommendations Assessments, Developments and Evaluation (GRADE) is a rating system developed by the GRADE Working Group in 2007,[20] shown in Table 5.

Integration With Clinical Expertise Into Practice

Arguably the most important aspect of EBM, or the goal of EBM if you will, is to integrate best evidence with clinical expertise for best treatment of a patient. The ability to integrate best evidence with clinical experience into practice is 2-fold: (1) one must be comfortable and capable in utilizing EBM in his or her practice, and (2) one must be able to understand and incorporate the patient's needs and wants to establish the best course to follow in terms of treatment and management.

When using the approach to practicing EBM discussed in this chapter, it is important to recall that the goal is to combine evidence, clinical experience, and patients' rights and perspectives to determine the solution. The importance of patients' perspectives, beliefs, expectations, and goals for life and health cannot be downplayed, because the approach to care in this EBM model is patient centered. By considering how patients think about the available options and their relative benefits, harms, costs, and inconveniences when determining options through evidence and clinical expertise, we engage in shared decision making. With this approach, we can make a compromise between these 3 factors to determine the best approach to a given patient in a given context.

TABLE 5. *GRADE Rating System*

Code	Quality of Evidence	Definition
A	High	Further research is very unlikely to change our confidence in the estimate of effect. ● Several high-quality studies with consistent results ● In special cases, 1 large, high-quality multicenter trial
B	Moderate	Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate. ● One high-quality study ● Several studies with some limitations
C	Low	Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate. ● One or more studies with severe limitations
D	Very low	Any estimate of effect is very uncertain. ● Expert opinion ● No direct research evidence ● One or more studies with very severe limitations

There have been various resources developed to help busy clinicians to identify and integrate the best available research evidence with their clinical expertise and the patient perspective discussed above. Clinical guidelines based on best evidence have been developed in a variety of specialties, providing a good starting place for commonly encountered scenarios.

Evaluation

An important component of practicing EBM is the fifth step: self-evaluation. After working through the EBM steps on a particular clinical question, it is important to review each of the 5 steps and evaluate whether it was completed in its entirety, effectively and efficiently. By continuously self-evaluating, gaps in a clinician's EBM skill set can be identified. A complete and helpful list of important questions to ask oneself in evaluation can be found in *Evidence-Based Medicine*: *How to Practice and Teach EBM*.[17]

- Are my questions specific and answerable? Am I able to form questions throughout the day and save time to target them later? Is the importance of asking good questions coming across in my teaching? Am I modeling this?
- Do I know the best resources for my questions? Do I have appropriate access to these resources? Am I searching from a wide variety of resources?
- Am I critically appraising the evidence I have found? Am I accurately and efficiently applying measures introduced here (likelihood ratio [LR], number needed to treat [NNT], relative risk reduction [RRR])?
- Can I work through particular concerns about management and integration to relate this evidence to a particular patient? Can I accurately and efficiently adjust my findings to fit my unique patient?

As mentioned earlier in this chapter, one of the important concepts that has fostered an environment where EBM can blossom is the idea of lifelong learning. Alongside self-evaluation, one of the most important techniques we can use to better ourselves as clinicians is to encourage and engage in continuing professional development. Developments in how we practice EBM, identified and updated through ongoing self-evaluation, are a part of this lifelong learning, while continually aiming to increase our knowledge base of the best evidence. What good is this evidence, however, without professional wisdom? Without professional wisdom obtained through ongoing professional development, evidence cannot be adapted to specific circumstances, and circumstances where evidence is not available would present quite a challenge.

CONCLUSIONS

This section has just but scraped the surface with regard to the impact of this paradigm shift in medical practice. This new approach to clinical decision making focused around the sound application of best research evidence is becoming so common in all fields of medicine that you would be hard pressed to find a physician or surgeon not familiar with RCTs, meta-analyses, Cochrane reviews, or evidence-based guidelines. As orthopaedics moves forward with the momentum of this global EBM movement, evidence-based orthopaedics is becoming a term, concept, and way of life in the clinical setting for all in the field. As discussed, it is not only the retrieval and appraisal of evidence that are important, but also how this evidence can be applied to a specific clinical situation considering societal values, as well as each patient's individual perspective. By learning how to approach searching for evidence in an effective and efficient manner and by learning where to look, how to look, and what you are looking for, the task of using evidence in everyday clinical practice becomes less and less daunting.

Lauren E. Roberts, M.Sc.
Jón Karlsson, M.D., Ph.D.
Mohit Bhandari, M.D., Ph.D., F.R.C.S.C.

SECTION 3

Levels of Evidence

Traditionally, surgical indications and treatment decisions have been based largely on expert opinion and personal experience. Although EBM has been proclaimed as one of the greatest achievements in internal medicine over the past 150 years,[21] its influence has been slow to seep into the surgical literature

because of the unique challenges of surgical trials. In 2003 levels of evidence were first introduced into *The Journal of Bone and Joint Surgery*, reflecting increased awareness of the importance of quality in an individual study.[22] This recognition of high-level research propelled the orthopaedic community to design and accomplish better studies,[23,24] which in other areas of medicine have ultimately led to significant treatment advances.[25]

Levels of evidence are important not only in determining whether a study is of higher quality than another, but they give the reader an immediate sense of how much weight the results of the study should be given.[26,27] The Oxford Centre for Evidence-Based Medicine has created a detailed hierarchy of evidence, in which the highest level remains a meta-analysis of homogeneous, high-quality RCTs.[28] A significant proportion of current orthopaedic studies are observational studies. To ensure standards of reporting observational studies, the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) statement was created, which assists investigators when reporting observational studies and supports editors and reviewers when evaluating these studies.[29] More recently, Grades of Recommendation Assessment, Development, and Evaluation (GRADEs) have been introduced to allow for a transparent and comprehensive method to grade the quality of evidence and strength of recommendations about the management of patients.[30]

HOW LEVELS OF EVIDENCE ARE ASSIGNED

What Is the Primary Research Question?

For a level of evidence to be assigned, one must first assess the primary research question. The level of evidence is assigned specifically to whether the primary research question, well-defined in the purpose section of a manuscript, was aptly addressed in the results and conclusion sections. Thus, asking a focused question helps yield a more answerable question, and assignation of level of evidence is relatively straightforward.

For example, in a study comparing the use of a bioabsorbable interference screw versus washer-post construct for tibial fixation in an anterior cruciate ligament (ACL) reconstruction, it would be ideal to only manipulate a single variable. In other words the study includes the same surgeon, same technique, and all patients with the same isolated ACL injury. The outcome would be a single data point, such as Lachman examination. In this way, the primary research question is focused on answering 1 specific question: "Does tibial fixation of the graft affect the postoperative Lachman examination?" If a difference in tibial translation is found between the 2 types of fixation, then a conclusion can be made as to whether or not there was a difference.

Conversely, it becomes very difficult to assign a level of evidence when the primary research question is not well-defined or the conclusions do not answer the research question. Frequently, studies will make conclusions based on their results, but in fact the conclusions were not related to the primary research question. Therefore it is extremely important when designing or reviewing a study to first evaluate whether the research question for the study is well defined and then evaluate whether the conclusions of the study are related to that primary research question.

Study Designs

Once the primary question is determined, the next task is to identify the study type. Levels of evidence can be divided into 4 different study designs: therapeutic, prognostic, diagnostic, and economic or decision analyses.[31]

Therapeutic Study Type: Therapeutic studies focus on assessing the effect of a specific treatment on the outcome of a specific disease process. A practical test to determine whether a study design is considered therapeutic is if the factor being studied can be allocated in a random fashion. For example, a study of ACL reconstruction evaluating the effect of graft type (e.g., bone-tendon-bone *v* hamstring autograft) on the outcome of reconstruction would be a therapeutic study because the graft type can be randomly allocated.

Prognostic Study Type: Prognostic studies evaluate the effect of patient characteristics on the outcome of a disease process. Prognostic studies differ from therapeutic studies because the factors being evaluated cannot be randomly allocated. For example, a study of the effect of age on outcome of ACL reconstruction in 2 different study groups (e.g., patients aged <30 years *v* patients aged >30 years) would be considered a prognostic study because age cannot be randomly allocated to 2 groups of patients in the study.

Diagnostic Study Type: Diagnostic studies are designed to assess whether a specific test is related to the presence or absence of a particular pathology. For

example, in patients with femoroacetabular impingement of the hip, the anterior impingement test can be performed for assessment. A study examining the effect of the anterior impingement test and its relationship to femoroacetabular impingement is an example of a diagnostic study design. Another example would be joint-line tenderness and its ability to detect meniscus tear.

Economic Analyses: Economic analyses are designed to assess the cost-effectiveness of a certain treatment for a certain pathology. For example, in the case of a group of young patients with femoroacetabular impingement, one might compare the cost-effectiveness of open versus arthroscopic impingement surgery.

Decision Analyses: Decision analysis studies are performed to evaluate the outcome of a certain therapy to determine the ideal treatment. For example, in evaluating surgical versus nonsurgical treatment for patients aged greater than 40 years with ACL deficiency, an expected-value decision analysis, which is a systematic tool for quantitating clinical decisions, can be used to conclude ACL surgical reconstruction as a preferred treatment.[32] An inherent limitation of this study type is that actual patients are not evaluated.

LEVELS OF EVIDENCE IN DETAIL

Several systems for rating levels of evidence are available.[33] The one chosen by *The Journal of Bone and Joint Surgery* and *Arthroscopy* has 5 levels of study design for each of 4 different study types: therapeutic, prognostic, diagnostic, and economic or decision modeling.[21,22,25] Among study designs, there exists a hierarchy of evidence, with RCTs at the top (Level I), controlled observational studies in the middle, and uncontrolled studies and opinion at the bottom (Level V).[33]

Understanding the association between study design and level of evidence is important. Higher levels of evidence should be more convincing to surgeons attempting to resolve clinical dilemmas.[21] Because randomized clinical trials are not always possible, Level I evidence may not be available for all clinical situations. Therefore Level III or IV evidence can still be of great value to the practicing orthopaedic surgeon. It is important to consider that an answer to a clinical question must be based on a composite assessment of all available evidence. No single study provides a definitive answer.

Level I

Therapeutic studies

1. RCTs with (*a*) significant difference or (*b*) no significant difference but narrow confidence intervals
2. Systematic reviews of Level I RCTs (studies were homogeneous)

Prognostic studies

1. Prospective studies
2. Systematic review of Level I studies

Diagnostic studies

1. Testing of previously developed diagnostic criteria in series of consecutive patients (with universally applied reference "gold" standard)
2. Systematic review of Level I studies

Economic and decision analyses studies

1. Clinically sensible costs and alternatives; values obtained from many studies; multiway sensitivity analyses
2. Systematic review of Level I studies

Level II

Therapeutic studies

1. Prospective cohort study
2. Lesser-quality RCT (e.g., <80% follow-up, no blinding, or improper randomization)
3. Systematic review of Level II studies or Level I studies with inconsistent results

Prognostic studies

1. Retrospective study
2. Untreated controls from an RCT
3. Systematic review of Level II studies

Diagnostic studies

1. Development of diagnostic criteria on basis of consecutive patients (with universally applied reference "gold" standard)
2. Systematic review of Level I and II studies

Economic and decision analyses studies

1. Clinically sensible costs and alternatives; values obtained from limited studies; multiway sensitivity analyses
2. Systematic review of Level II studies

Level III

Therapeutic studies

1. Case-control study

2. Retrospective cohort study
3. Systematic review of Level III studies

Diagnostic studies

1. Study of nonconsecutive patients (without consistently applied reference "gold" standard)
2. Systematic review of Level III studies

Economic and decision analyses studies

1. Analyses based on limited alternatives and costs; poor estimates
2. Systematic review of Level III studies

Level IV

Therapeutic studies
Case series (no, or historical, control group)
Prognostic studies
Case series
Diagnostic studies

1. Case-control study
2. Poor reference standard

Economic and decision analyses studies
No sensitivity analyses

Level V

Therapeutic studies
Expert opinion
Prognostic studies
Expert opinion
Diagnostic studies
Expert opinion
Economic and decision analyses studies
Expert opinion

EXAMPLES OF STUDIES OF DIFFERENT LEVELS OF EVIDENCE

Level I

In a study of a consecutive series of patients with the diagnosis of internal snapping hip syndrome, patients were randomized into 2 different methods of endoscopic release of the iliopsoas tendon.[34] Patients in group 1 were treated with endoscopic iliopsoas tendon release at the lesser trochanter, and patients in group 2 were treated with endoscopic trans-scapular psoas release from the peripheral compartment. A quality randomization process included randomizing patients at the last possible time point, e.g., at the time of surgery. An a priori power analysis was performed to ensure adequate numbers of patients in each randomized group. Preoperative and postoperative clinical and imaging assessments were evaluated for all patients. No statistical difference was found between groups. Therefore this RCT, with a defined and appropriate sample size and narrow confidence intervals, is characterized as a Level I study, even though no statistically significant difference was determined.[34]

Level II

In a prospective cohort study, patients aged older than 40 years were compared with a group of patients aged younger than 40 years who underwent autologous chondrocyte implantation for isolated cartilage defects of the knee.[35] The authors' hypothesis was that the older group of patients would have inferior clinical results compared with the younger group of patients. All patients were followed up for 2 years, and validated clinical outcomes were used. The authors' hypothesis was disproved, because there was no statistically significant difference in the 2 groups of patients treated with autologous chondrocyte implantation. This prospective study does not obtain a Level I designation because it is nonrandomized.

Level III

The efficacy of open versus arthroscopic Bankart repair remains controversial; therefore the authors designed a retrospective case-control study to determine whether there is a significant difference in cost between the 2 surgical procedures. In a Level III retrospective case-control study, the authors retrospectively reviewed the medical records and billing information of consecutive patients treated for recurrent, post-traumatic anterior shoulder instability.[36] They compared 22 patients who had open Bankart repair with 20 patients who had arthroscopic Bankart repair. Total operating times and all charges were obtained from records. Patients were also clinically evaluated. This study found similar shoulder scores and rates of dislocation between the 2 groups. The arthroscopic Bankart repair had a lower cost, but if an obligatory overnight inpatient stay was taken into account, the cost difference was negligible. Because of its retrospective nature, this study was characterized as a Level III, therapeutic cohort study.[36]

Level IV

The purpose of a Level IV study is to retrospectively review the outcome of a group of patients

treated in a similar way. In a Level IV therapeutic case series study, the authors described transphyseal ACL reconstruction with hamstrings performed in 26 patients with open tibial and femoral physes.[37] Clinical and radiologic outcomes were evaluated retrospectively. Their outcomes were well defined, with validated knee scores and detection of any growth disturbance on scanograms. They concluded that their technique yielded good outcomes and no growth disturbances. Because the authors did not compare their technique with another technique, and given its retrospective nature, this represents a Level IV therapeutic case series.

Level V

In a Level V study, the authors showed the use of the 70° arthroscope for several arthroscopic procedures and in a number of circumstances in which it offers superior visualization to a 30° arthroscope.[38] In this study the authors demonstrated their particular expertise with this arthroscopic instrument, which may be interesting for arthroscopic surgeons who are not familiar with the 70° arthroscope. However, because this study does not report any results or clinical outcomes, it is considered expert opinion.

GRADES OF RECOMMENDATION

As surgeons, we often find multiple studies to be more convincing than a single article. Although the appropriate literature on a particular clinical question can be identified in many ways, to search the literature ourselves is time-consuming and the search may not be comprehensive. Although review articles are often comprehensive in the available evidence they include, the conclusions that they contain can be uncertain. Therefore grades of recommendation have been introduced in the development of practice guidelines. In this process a reviewer or organization can gather all the appropriate literature, appraise the literature by assigning a level of evidence, and summarize the overall quality by allocating a grade of recommendation.[26,30] This helps the reader by giving definitive treatment recommendations that should definitely (grade A) or probably (grade B) guide treatment decisions for their patients. In addition, a grade of I, or insufficient or conflicting evidence not allowing a recommendation for or against intervention, advises a surgeon to treat patients based on his or her best judgment or on a case-by-case basis.

Grade A indicates good evidence (Level I studies with consistent findings) for or against recommending intervention. Grade B indicates fair evidence (Level II or III studies with consistent findings) for or against recommending intervention. Grade C indicates poor-quality evidence (Level IV or V studies with consistent findings) for or against recommending intervention. Grade I indicates that there is insufficient or conflicting evidence not allowing a recommendation for or against intervention.

CONCLUSIONS

The purpose of this chapter is to provide the orthopaedic sports medicine surgeon with a better understanding of the levels of evidence and their clinical implications. Such understanding is extremely helpful, not only from a research design standpoint but also to aid readers in understanding the importance of a particular study's conclusions.

From a design standpoint, in order for a research protocol to maximize the best possible level of evidence, it is important for the orthopaedic sports medicine researcher to consider levels of evidence when outlining the primary research question. Furthermore, it is important to recognize that performing a Level I surgical study is extremely difficult, because it requires a significant amount of preparation, time, and financial investment to allocate resources. Level II, III, and IV studies have their own worth and merit and are especially useful in the circumstances where Level I studies would not be feasible. When observational studies are being performed, the STROBE recommendations will assist the investigator in maintaining methodologic transparency and also assist the reader in comprehensively evaluating the quality of the study.

From a reader's standpoint, if a study is assigned Level I evidence, and a grade A recommendation, then the reader can feel confident that the results of the study have the highest level of validity. In this situation the reader/surgeon may choose to change clinical practice based on those recommendations, thus shaping and directing future sports medicine care. Ultimately, it is this endpoint, the best care for patients, that is our highest goal.

Aaron J. Krych, M.D.
Bruce A. Levy, M.D.
Mario Ferretti, M.D., Ph.D.

SECTION 4

Study Designs: Randomized Trials, Level I Evidence, CONSORT Checklist

The randomized clinical/controlled trial (RCT) represents the highest level of evidence or study design. In orthopaedic surgery and sport medicine, there are a multitude of questions requiring evidence-based answers. It will become increasingly more important to perform RCTs to address these questions. This chapter identifies the problems encountered by the surgeon, as well as the strategies and how to address these concerns.

WHY PERFORM RCTS?

An RCT is the most valid study design to evaluate the efficacy or effectiveness of surgical treatments. Efficacy refers to the ideal situation with optimal patient selection, well-controlled surgical technique, postoperative compliance, and so on. This type of randomized trial is sometimes referred to as an explanatory trial.[39] Effectiveness refers to a more real-world situation, where the patients have more variability in their disease state, multiple surgeons may be involved, and the postsurgical course is less well-controlled and more typical of most surgeons' practices. This type of RCT is sometimes referred to as a pragmatic trial.[30]

The RCT is prospective by definition, and therefore this term is redundant. All RCTs are prospective because the primary research question being addressed, independent variable (treatment), dependent variable (outcome), and inclusion and exclusion criteria should all be determined a priori. Patient recruitment and enrollment, consenting, randomization, data collection, and analysis are subsequently performed in a forward direction. Patients are randomly allocated to different treatment groups and are typically followed up in an identical manner with the main outcome of interest measured at a specific period of time. The groups of patients are similar with respect to known characteristics (e.g., inclusion and exclusion criteria) and unknown characteristics (those present by chance). Provided that an appropriate sample size is calculated and recruitment is achieved, the unknown characteristics are likely to be equally distributed between groups. Therefore, if a difference in outcome is identified, the findings can be directly attributed to the efficacy or effectiveness of the specific surgical treatment.

Therefore the RCT is less likely to introduce bias because treatment group assignment is randomly determined. Other biases can occur, however, despite the randomized design. These would include a lack of blinding of the patients or assessor, different follow-up times, loss to follow-up, differential exclusions, expertise-based bias, and early reporting of results before the full sample size is achieved.[40] Whatever bias is introduced must first be recognized, and then it may be accounted for.

So, if we truly want to determine the benefits of one operative technique over another, surgical versus nonsurgical treatment, different rehabilitation protocols, and so on, the RCT is the best possible study design. We must be cognizant of the fact that the design is only one component of conducting valuable research. Randomization does not compensate for poor adherence to all other important methodologic issues.

REASONS FOR NOT PERFORMING AN RCT

The Problem

Randomized clinical trials are only necessary if the clinical problem/surgical treatment is common; if the question is a significant issue to clinicians and patients; and most importantly, if the answer to the clinical question is clearly not known. It would be unnecessary to perform a randomized clinical trial in circumstances where observational studies are so compelling and/or outcomes are dramatic and life-saving (e.g., amputation compared with antibiotic treatment for clostridial gas gangrene of the foot). In other words, the treatment effect is so large, and the consequences so grave, that it is not necessary to compare the surgical procedure with existing treatment. The same can be said for parachute use compared with placebo.[41] All problems are not amenable to an RCT.[42] Until the rules of surgical engagement change to be similar to those related to medical ther-

apy it is unlikely that surgical RCTs will become the norm rather than in the minority.[43]

The Patient

Patients present with preconceived notions about what treatment is best. Whether they have talked to a friend or family member, someone in health care, a physician, or even another surgeon, there is a bias that each patient may have. Patients will typically investigate their problem on the Internet. They will identify specific operations and other surgeons who have addressed their problem with a particular procedure.

There are different cultural expectations around the world. At the same time that an RCT comparing outpatient with inpatient ACL surgery was being performed in Canada, patients were staying in the hospital for 2 weeks or more in Europe and Japan.[44]

Patients may simply want to know what procedures they will be undergoing, and any doubt leads to a lack of confidence. Some patients will consent to a trial because it is the only chance that they can undergo the latest procedure (i.e., being included in the experimental group rather than undergoing the usual technique). Some patients feel the exact opposite sentiment.

There is a permanency regarding surgery. This can affect a patient's decision, such as in a trial comparing surgical with nonsurgical treatment. If a patient ends up in the nonsurgical arm, there is a perception that if the treatment fails, then surgery may be an option. However, once a procedure is performed, there is no going back. These patient-related concerns influence whether eligible patients are open to recruitment into a trial.

The Performer (Surgeon)

The surgeon may be the greatest barrier to surgical trials! Compared with medical therapies where there are strict regulations on how a drug is released, surgical innovation can occur with little or no restraint from regulating agencies, hospitals, or local ethics committees. This is definitely the case when there is a minor variation in technique or implant used. Therefore there is no incentive whatsoever to perform a trial to determine the efficacy of a particular procedure.

Surgeons are innovators. Arthroscopic surgery may not have become the gold standard if we had the requirement of randomized clinical trials to show its benefit. Historically, arthroscopy of the knee was more expensive, took longer to perform compared with the equivalent open procedure, and was fraught with complications.

The irony of being a surgeon is that we can perform "experimental" surgery on our patients with their consent and with little or no regulation, but if we want to perform an experiment (i.e., an RCT), then we must obtain both scientific and ethical approval.[45]

The Procedure

There are many barriers to performing an RCT regarding the surgical procedure. It is well-recognized that there is a learning curve with respect to any operation. One could argue that, if it is a minimal change, then the learning curve is shallow but a significant departure from what is usually done may in fact have a very steep curve. A certain minimum standardization is required in any trial, particularly if there is more than 1 surgeon involved to ensure consistency between surgeons. If only 1 surgeon is involved, then it may be the case that he or she is better at 1 procedure compared with another. We can take the analogy from the sport of tennis. It is well known that only a few players in history have been able to master the game on hard courts, clay, and grass all in the same year. Why would we expect surgeons to be able to perform different procedures (e.g., arthroscopic compared with open) equally as well?[46] Therefore, if more surgeons are required (to increase sample size), then strategies such as training manuals, stratification by surgeon, and matching surgeon experience are techniques that can alleviate the variability of the procedure. One recently applied method has been called the "expertise-based design."[40] In this RCT the patient is randomized to the procedure, and the surgeon with the expertise for that particular procedure carries out the surgery.[40,47]

The Process

This is probably what scares most surgeons away from performing a randomized trial: the process from start to finish is overwhelming to most surgeons. Statements include the following: It is going to take too much time! We do not have the resources! We will never get it through ethics! We do not have enough patients! I get good results with the technique I am familiar with! We do not need to do a trial; I can do a bunch of these new procedures and compare to what I have done in the past! I am too busy taking care of patients to do research! I want to do the latest techniques; it is what my patients expect! It is not my responsibility to do research; let the researchers figure this out! I do not have a research assistant! It takes too much money!

There is no doubt that conducting a randomized trial requires significant infrastructure support, time, and effort. The process is daunting at first and difficult to

implement but eventually routine. Until there are stricter rules regarding the use of new procedures, surgeons will not be compelled to be involved in appropriate trials.[45]

HOW DO WE SOLVE THESE PROBLEMS AND WHAT ARE THE PREREQUISITES?

It is much easier to identify reasons not to do something. It is easier for a medical student to answer a difficult question with a negative response. For example, when asked about the differential diagnosis of an acute hemarthrosis in the knee, he or she may say, "It is not a tumor." Although this statement is correct and easy to identify on the differential, it is not a very useful answer in evaluating the patient. The solution lies in the following concepts, the 3 C's of successful trial research: clinical equipoise, commitment, and collaboration.

Clinical Equipoise

Clinical equipoise is defined as genuine uncertainty over whether one treatment or another is beneficial. This equipoise should involve the expert clinical community where there is "honest professional disagreement among . . . clinicians."[48] Therefore, if one or more treatment options are available for a particular problem and the best option is not known, then we have reason to consider a randomized clinical trial. It is necessary to build a case for clinical equipoise, which essentially is the essence of the rationale for the RCT. A surgeon should first review the available evidence in a systematic way, analyze the results (with a meta-analysis if possible), and determine an answer to his or her question. If this answer is clear, then there is no equipoise and, therefore, no need to perform another trial.

However, clinical equipoise relates to not only uncertainty regarding treatment options but the ethics of performing a trial from all perspectives: the patient's, the surgeon's, and society's. It is necessary to ask the question, Who is uncertain, the individual surgeon or the community of surgeons? A surgeon may consider randomizing his or her patients to a particular arthroscopic fixation technique such as absorbable compared with nonabsorbable suture anchors. Whereas this trial may be easy to perform because there is little impact on the patient's decision making, it may not matter to the community of surgeons or society as a whole. Therefore, is it really worth performing a randomized clinical trial? Patients just like surgeons are influenced by their position on the equipoise spectrum. They may desire to understand in great depth the treatment options, they may want to appreciate the bigger-picture perspective of helping medical science and therefore the surgical community perspective, or they may in fact simply trust the surgeon.

Clinical equipoise requires not only the consideration of the individual surgeon's perspective but the community of surgeons that establishes standards of practice. Most surgeons have difficulty with this concept, and therefore failure of consensus of evidence within the clinical community is the usual driver for a trial. Ultimately, the ethical surgeon must do what is best for his or her individual patient. Uncertainty is a moral prerequisite for being involved in an RCT, but if we know the best and correct treatment, then we should perform it.[45,48,49]

The following example should illustrate the concept of clinical equipoise and the moral or ethical responsibility of the surgeon. A randomized clinical trial was conducted to compare 3 surgical techniques for ACL reconstruction.[50] A meta-analysis had been conducted to determine whether an autograft patellar tendon compared with autograft hamstring reconstruction resulted in improved outcomes for patients at 2 years. Not only was the meta-analysis inconclusive but it identified many concerns with respect to the available evidence.[51] At the same time, surgeons were advocating the so-called double-bundle technique for ACL reconstruction. Therefore it seemed logical to conduct a randomized clinical trial comparing the existing techniques with the newer double-bundle procedure. This would represent established clinical equipoise. However, there was a clinically identifiable subgroup of patients who have the prerequisite diagnosis of an ACL-deficient knee but whose knees on careful examination had minimal documentable translational and rotational instability. These patients, also on careful arthroscopic examination, had identifiable ACL tissue that was both biologically viable and mechanically supportive. Whether this represents a partially torn ACL, a single-bundle ACL failure, or healing of a complete ACL tear is debatable. However, the surgeon believed on moral and ethical grounds that these patients should be excluded from the trial. This was based on the principle of biological preservation of the patient's own tissue and, most importantly, the empirical evidence from his own practice that this subgroup of patients had a better outcome than those who had undergone reconstruction in the usual way. This example demonstrates the difficulty with addressing both clinical equipoise and the ethics of performing a randomized clinical trial.

Commitment

Probably the most important prerequisite for conducting an RCT is one's commitment. Commitment relates to not only being involved with respect to one's role but more importantly being committed to the question rather than finding the answer. This is particularly difficult for surgeons because we are driven to solve patients' problems through our own individual surgical skills.[30,42,43,45,52-56] Surgeons are typically characterized as innovative rather than reflective, performers rather than observers, and are interested in immediate gratification rather than long-term rewards. The RCT requires a different type of commitment that is reflective before and after the fact, requires persistence, and may lead to an answer that is not consistent with what is expected. For example, in a trial comparing surgical with nonsurgical treatment, it is inherently difficult for a surgeon to be committed to the question unless there is a perceived problem with the surgical option. Our training, experience, and rewards are derived from the outcomes of the patient's surgical treatment.

As innovators, surgeons become readily aware of the latest technique or improvements on previous procedures. It might take 2 to 5 years to recruit enough patients into a trial and then the requirement of a minimum 2-year follow-up. The process of writing the proposal, obtaining funding, and obtaining ethical approval may take at least 1 year. A meaningful trial may take anywhere from 5 to 10 years to complete. During this time, the surgical world has moved forward, techniques have been modified, case reports may have suggested complications with a particular procedure, and so on.[57]

The surgeon must therefore act and be committed in a way that is somewhat foreign to his or her normal existence. This commitment is compounded by the fact that in every respect conducting a surgical trial takes more time and effort than what is necessary to run a clinical practice. Successful surgical "trialists" are just as passionate about the research as any aspect of their clinical practice. They likely spent additional time learning how to perform research in the fields of clinical epidemiology, methodology, or public health in addition to their clinical fellowships.

Collaboration

We are not aware of any surgeon in the world who is a methodologist and biostatistician, has extra time to devote to clinical research, and also a large enough clinical practice to conduct a meaningful randomized clinical trial without help.

The collaborative infrastructure support is not only helpful but necessary. One solution is to pay for collaborative support by hiring a research organization to conduct the trial and therefore enter patients and perform the surgery. Another approach is to identify individuals with the expertise in methodology and biostatistics who will be partners/co-authors in the trial and therefore will provide their expertise without financial compensation. There will always need to be a research coordinator or assistant. This individual provides day-to-day support and keeps the trial moving forward, addressing all of the details and complexities of conducting an RCT.

Collaboration may take the form of including other clinicians and surgeons. These individuals may be at the same institution or could be based out of multiple centers. In these circumstances the clinicians will need to have the same requisite ethical and clinical equipoise to the primary surgeon and the time and commitment necessary for success. It is well-recognized in multicenter trials that the host site is usually more successful in recruiting patients and conducting all aspects of the trial.[58,59] The exception to this is when the trial is funded centrally and the collaborating centers have financial incentives to recruit and follow up the patients.

CONDUCTING A RANDOMIZED CLINICAL TRIAL?

Once the prerequisites have been addressed (i.e., an important clinical concern, commitment to the question, and collaboration), the trial is ready to be implemented.

However, implicit within these prerequisites is that the research question has been carefully refined and a detailed proposal drafted, reviewed, and rewritten, along with an application for funding and ethical approval.[58,60] The implementation starts once approval and funding have been achieved.

It is necessary to engage all people and settings (hospital wards [i.e., emergency, inpatient, and outpatient], clinics, operating rooms, and so on) that may or may not be impacted by the trial. This process should ideally occur during the proposal stage but is an obligatory part of implementation. Informing and engaging everyone must occur before, during, and after conducting the trial if it is to be successful. Simple incentives such as providing refreshments to the hospital or clinic staff or giving gift vouchers to referring physi-

cians have proven to be cost-effective ways to facilitate this critical engagement.

Informing the medical community of the trial is also very important at the start. This includes presenting at rounds and business meetings, advertising the trial, and registering the trial in an international database.

Within the written proposal are specific criteria on how the patient population is to be sampled. When patients are seen, they need to be screened to determine whether they fit the inclusion and exclusion criteria. Assuming the patient is eligible, the consenting process can proceed. Informed consent is a critical time-consuming activity. There may be ethical issues with respect to who obtains consent, and this is typically regulated through each individual institution. It has been our experience with surgical trials that a surgeon is the person best suited to obtain consent for a surgical randomized clinical trial. This leads to higher recruitment of patients.

The process of randomization can take many forms, and there are different types of randomization. With respect to surgical trials with more than 1 surgeon involved, stratification by surgeon is necessary unless 2 surgeons are matched for all known characteristics such as experience, location, and so on.

One technique to help the process of randomization is called expertise-based randomization.[40] This is where the patient is randomized to the procedure before going to the operating room. This technique provides the surgeon the ability to participate in an RCT but still retain his or her independence and individual preference to perform his or her procedure of choice. We have used this expertise-based randomization technique successfully when comparing open and arthroscopic procedures in the shoulder.[46]

Irrespective of the type of randomization, there are specific requirements that must be adhered to. These include allocation concealment and adequate sequence generation, i.e., typically, computer-generated random-number sequencing.[39] Although opaque envelopes are considered an appropriate concealment technique, they can be tampered with and the sequence identified. Current standards would involve a Web-based remotely accessed computer-generated randomization process that is performed by someone independent of the surgeon or primary investigator.[39]

The randomized trial should include a primary outcome (i.e., the dependent variable), such as a validated patient-reported outcome, and the defined intervention (i.e., the independent variable), such as the standard operation compared with the new surgical procedure.

The sample size for the RCT is directly related to the primary outcome, the measured treatment effect (i.e., the expected clinical difference between the 2 treatment groups), and the variability of the outcome measured in the standard deviation. Ideally, the expected difference and variability of the outcome are known values based on previous pilot data or data from similar populations of patients.[39,53,54] Without this information, the sample size calculation becomes speculative, and therefore the trial may be underpowered to show a meaningful clinical difference between treatment groups. In general, the more precise (i.e., less variability) the outcome and the greater the expected differences between treatment groups, the smaller the sample size. In addition, those dependent variables that are measured on a scale that allows for correct statistical analysis with means and standard deviations (i.e., parametric statistics) are likely to require a smaller sample size. Trials where the primary outcome is a probability (i.e., nonparametric statistics) are more likely to require a greater sample size.

The greatest barrier to conducting a surgical trial is recruitment and therefore meeting the a priori sample size.[59] Surgeons typically overestimate the number of patients who would be eligible, and the eligible patients do not always consent to the trial.[57] Some of the strategies to improve recruitment include collaborating with more surgeons, involving multiple centers, using baseline data to recalculate the sample size (assuming that there is less variability), providing incentives to include patients, continual strategies to engage people, ensuring regular patient contact to avoid loss to follow-up, and modification of inclusion and exclusion criteria to be more inclusive with respect to eligibility.

Once all of the details of the trial are organized, carrying out the trial is arguably the easiest part. It necessitates the help of the coordinator and assistants, and it requires a time commitment; however, as people in the clinics, wards, and operating rooms become familiar with the process, the trial should move ahead smoothly.

Every strategy to maintain contact with the patients should be used. This may include regularly scheduled follow-up visits, phone communication, and use of e-mail or social media.

Once the data are collected and the patients have been followed up, the analysis will occur. The help of a biostatistician is usually necessary for randomized trials.

The results will be interpreted, presented, and subsequently published.

REPORTING RCTS: THE CONSORT CHECKLIST

The randomized clinical trial represented the gold standard for evaluating interventions, but the accuracy of such trials' reporting was not consistent and therefore bias could be introduced. A worldwide group of researchers, methodologists, and clinicians, concerned that the reporting of trials lacked lucid and complete descriptions of the critical information, created the Consolidated Standards of Reporting Trials (CONSORT) statement (1996).[61,62] This has undergone recent revision, in 2010.[63,64] The CONSORT 2010 statement, checklist, and flow diagram provide authors with guidance on how to report their trials. The flow diagram (Fig 3) illustrates the progress of the trial from the start and includes the following: (1) the enrollment phase with numbers of eligible patients and those excluded for reasons such as not meeting inclusion criteria or declining to participate or for other reasons, and the number of patients randomized; (2) the allocation phase, which includes the exact numbers of patients who were allocated to the treatment groups, whether they received the allocated treatment, and if not, why not; (3) the follow-up phase, which includes the numbers lost to follow-up and the reasons why; and (4) the analysis phase, which includes those patients in each group who were analyzed and any who were excluded and for what reasons.[63,64]

The checklist (Table 6) represents a much more detailed list of characteristics of the trial that need to be reported.[63,64] The list includes the title and structured abstract, an introduction, the methods, the results, a discussion, and a section for other information. The checklist requires the authors to identify within their manuscript the page number where the appropriate information is written. Important concepts include the background and objectives, the trial design, the patients, detailed descriptions of the interventions, whether the outcomes were completely defined and prespecified, sample size determination, blinding of

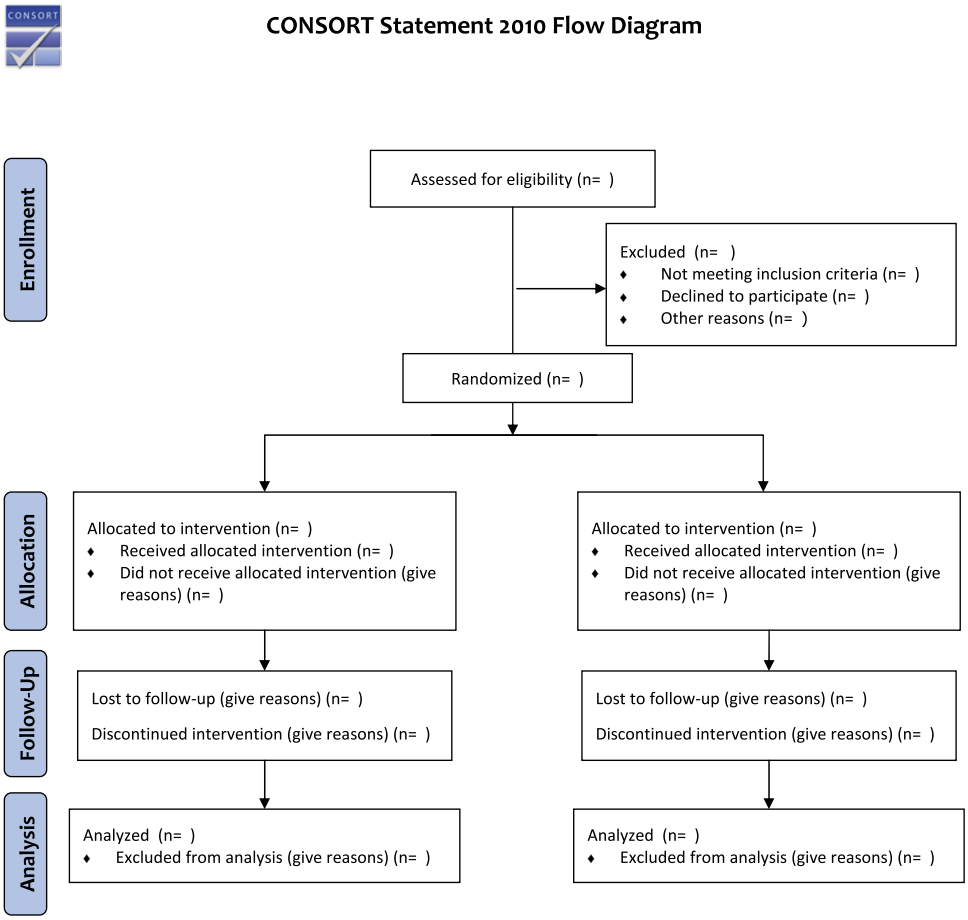

FIGURE 3. CONSORT flowchart.

TABLE 6. *CONSORT 2010 Checklist of Information to Include When Reporting a Randomized Trial**

Section/Topic	Item No	Checklist item	Reported on page No
Title and abstract			
	1a	Identification as a randomised trial in the title	
	1b	Structured summary of trial design, methods, results, and conclusions (for specific guidance see CONSORT for abstracts)	
Introduction			
Background and objectives	2a	Scientific background and explanation of rationale	
	2b	Specific objectives or hypotheses	
Methods			
Trial design	3a	Description of trial design (such as parallel, factorial) including allocation ratio	
	3b	Important changes to methods after trial commencement (such as eligibility criteria), with reasons	
Participants	4a	Eligibility criteria for participants	
	4b	Settings and locations where the data were collected	
Interventions	5	The interventions for each group with sufficient details to allow replication, including how and when they were actually administered	
Outcomes	6a	Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed	
	6b	Any changes to trial outcomes after the trial commenced, with reasons	
Sample size	7a	How sample size was determined	
	7b	When applicable, explanation of any interim analyses and stopping guidelines	
Randomisation:			
Sequence generation	8a	Method used to generate the random allocation sequence	
	8b	Type of randomisation; details of any restriction (such as blocking and block size)	
Allocation concealment mechanism	9	Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the sequence until interventions were assigned	
Implementation	10	Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions	
Blinding	11a	If done, who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes) and how	
	11b	If relevant, description of the similarity of interventions	
Statistical methods	12a	Statistical methods used to compare groups for primary and secondary outcomes	
	12b	Methods for additional analyses, such as subgroup analyses and adjusted analyses	
Results			
Participant flow (a diagram is strongly recommended)	13a	For each group, the numbers of participants who were randomly assigned, received intended treatment, and were analysed for the primary outcome	
	13b	For each group, losses and exclusions after randomisation, together with reasons	
Recruitment	14a	Dates defining the periods of recruitment and follow-up	
	14b	Why the trial ended or was stopped	
Baseline data	15	A table showing baseline demographic and clinical characteristics for each group	
Numbers analysed	16	For each group, number of participants (denominator) included in each analysis and whether the analysis was by original assigned groups	
Outcomes and estimation	17a	For each primary and secondary outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval)	
	17b	For binary outcomes, presentation of both absolute and relative effect sizes is recommended	
Ancillary analyses	18	Results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing pre-specified from exploratory	
Harms	19	All important harms or unintended effects in each group (for specific guidance see CONSORT for harms)	
Discussion			
Limitations	20	Trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses	
Generalisability	21	Generalisability (external validity, applicability) of the trial findings	
Interpretation	22	Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence	
Other information			
Registration	23	Registration number and name of trial registry	
Protocol	24	Where the full trial protocol can be accessed, if available	
Funding	25	Sources of funding and other support (such as supply of drugs), role of funders	

*We strongly recommend reading this statement in conjunction with the CONSORT 2010 Explanation and Elaboration for important clarifications on all the items. If relevant, we also recommend reading CONSORT extensions for cluster randomised trials, non-inferiority and equivalence trials, non-pharmacological treatments, herbal interventions, and pragmatic trials. Additional extensions are forthcoming: for those and for up to date references relevant to this checklist, see www.consort-statement.org. Reprinted with permission.[63]

patients and investigators, the specifics of randomization, and details regarding the analysis and results, along with a discussion about the limitations and generalizability of the trial.

LIMITATIONS OF RCTS

Randomized clinical/controlled trials are really not suited to clinical problems that are uncommon or unimportant. The following examples demonstrate the limitations of the RCT design.

Example 1: To determine whether prophylactic antibiotics are required for routine arthroscopic meniscectomy of the knee in an otherwise healthy patient would be ludicrous. If the reported rate of infection is 3:1,000 patients, then to reduce the infection rate to 2:1,000 would require a sample size of nearly 40,000 patients per group.

Example 2: Several companies have developed new and innovative treatments for chondral lesions in the knee. These treatments have typically been evaluated in animal models with promising results. However, the animal model is not likely bipedal and probably uses a knee joint that is otherwise uninjured, normally aligned and the lesion is surgically prepared in one femoral condyle only. Subsequent human randomized trials would require a patient population that has an isolated chondral lesion to one femoral condyle in an otherwise stable and normally aligned knee that has failed standard treatment. In fact, these patients are very difficult to find. Assuming that the trial is completed, then the inferences from this trial and the results can only be applied to patients with similar characteristics to the original limited numbers of patients included in the trial.

Example 3: A randomized clinical trial comparing electrothermal arthroscopic capsulorrhaphy (i.e., the heat probe) versus open inferior capsular shift in patients with primary capsular redundancy was carried out in Canada.[57] This trial was hampered by several anecdotal reports of complications associated with the use of the heat probe. The trial was recently completed, and patient recruitment was the largest issue. The trial took 10 years to complete, and upon completion, the electrothermal arthroscopic capsulorrhaphy technique had been all but abandoned.

Example 4: In this hypothetical example, an RCT reports that surgical treatment is better than nonsurgical treatment but it comes to light that only 30% of the eligible patients were included in the trial. The other 70% of eligible patients may differ in several important characteristics. If this population is not accounted for, or their demographics are not compared with the trial patients, then the results may be very biased. If it turns out that there are known prognostic factors relating to patient outcome and these are dramatically different between the 2 populations, then the results of the trial are very limited.

Randomized trials are typically limited in that the strict nature of performing a trial requires specific inclusions and exclusions; consenting patients; sites with the necessary infrastructure; financial support, which may be from industry; and so on. The obvious conclusion asks the question of whether or not the results are generalizable to an individual orthopaedic surgeon.

IMPACT OF AN RCT: DOES IT CHANGE CLINICAL PRACTICE?

One of the largest problems of EBM and specifically conducting and reporting randomized clinical trials is that of knowledge translation. Peer-review funding agencies have developed strategies to ensure that the information gets to the end user, whether this is the patient or surgeon. One strategy is to provide specific funds within the grant for this purpose alone. Investigators applying for funding are therefore obligated to provide their strategies to disseminate the information. Different journals have partnered with funding agencies or organizations to provide a vehicle for authors to publish their results. Some agencies provide prizes for the best research in order to improve knowledge translation.

However, if the trial has internal and external validity and is reported widely, then there is every expectation that the results will have an impact on clinical practice. An example is the trend toward functional nonsurgical treatment of Achilles tendon ruptures based on recently published RCTs.[65] Ironically, rapid change in clinical practice is much more likely to occur if serious adverse events are reported from case series or through database surveillance rather than an RCT.[66]

CONCLUSIONS

Randomized clinical (controlled) trials in orthopaedic surgery represent the minority of published studies. The RCT is the most valid design to address clinically important questions. The question to be answered must be commonly encountered, must be important, and must show clinical equipoise. The surgeon must be committed, collaborate, and follow the necessary steps to perform a successful RCT.

Nicholas Mohtadi, M.D., M.Sc., F.R.C.S.C.
Mohit Bhandari, M.D., Ph.D., F.R.C.S.C.

SECTION 5

Study Designs: Cohort Studies, Level II Evidence, STROBE Checklist

Cohort studies allow investigators to examine a given population over a period of time for events related to the health of individuals in that population. Whereas case-control and cross-sectional studies evaluate patients at a single point in time, the defining characteristic of a cohort study is that it follows a group over time. In a typical scenario, investigators of a cohort study will collect health-related information at the beginning of the time period in question, or "baseline," and at predefined time points thereafter. The goal is to uncover, or assess the validity of, relations between a health-related variable recorded at one point in the study and an outcome recorded later. For example, all patients undergoing total knee arthroplasty at a given site are asked to fill out a questionnaire that, among other things, asks them whether they have diabetes. At the 10-year time point, some patients will be enjoying mobility whereas others may have a poor outcome. On noticing a pattern in their own patients, investigators might wish to evaluate the hypothesis that total knee arthroplasty patients with diabetes have worse outcomes than those without diabetes. Comparison of the outcomes of diabetic patients with nondiabetic patients would provide useful evidence to support or refute that hypothesis.

This example points to another important aspect of any cohort study design: the selection of the cohort. A cohort is defined by some shared characteristic. In the example above, the characteristic was that the patients underwent total knee arthroplasty. Completion of that procedure is an "eligibility criterion" for membership in the cohort. There is a broad distinction in cohort study design between a "closed cohort" and an "open cohort." A closed cohort is fixed; that is, the group is chosen at the beginning of the time period with which the study is concerned and does not change over the course of the study. The example above would be a closed cohort study if it were determined at the outset that exactly 100 patients would be included, all having undergone total knee arthroplasty at the given site during the month of July, 2010. An open cohort study follows a group whose membership changes over time. If the investigators of the arthroplasty study above aimed to continue to enroll patients indefinitely, the cohort's composition could change over time.

The STROBE checklist (Table 7)[67] was created to improve the quality of reporting of observational research, including cohort studies. It includes 22 items grouped according to the generally accepted components of a scientific article. This chapter will focus on those checklist items that are uniquely applicable to cohort study design: description of eligibility criteria and matching criteria (where applicable), explanation of loss to follow-up, summarized reporting of follow-up times, and quantitative reporting of outcome events.

STROBE: TITLE, ABSTRACT, AND INTRODUCTION

The title of any observational study should include a term denoting the design of that study, and cohort studies are no exception. An appropriate title might be "Incidence of Osteoarthritis After Meniscectomy: A Cohort Study." Not only does this allow the reader to quickly ascertain an important characteristic of the study—its design—but it provides for more effective electronic database searching because the study design can be indexed by its title.

The abstract describing a study must adhere to varying requirements set forth by individual journals, but there are certain items that should be addressed regardless of the format in which they are presented. These include, for one, background information; the impetus for the study should be explained. Authors should also state the specific objective(s) of the study and restate the study's design as indicated in the title. The setting of the study and details of patient selection, such as matching methods, are essential, in addition to the particulars of any measurements made. Naturally, authors should briefly report results and the conclusions they draw from those measurements, along with any inherent limitations of the study. A good abstract is both concise and comprehensive.

The introduction should elaborate on the context and objectives of the study as stated in the abstract. Authors should provide an overview of what is known

TABLE 7. *STROBE Statement: Checklist of Items That Should Be Included in Reports of Observational Studies*[67]

	Item No.	Recommendation
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract (*b*) Provide in the abstract an informative and balanced summary of what was done and what was found
Introduction		
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported
Objectives	3	State specific objectives, including any prespecified hypotheses
Methods		
Study design	4	Present key elements of study design early in the paper
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection
Participants	6	(*a*) Cohort study—Give the eligibility criteria, as well as the sources and methods of selection of participants; describe methods of follow-up Case-control study—Give the eligibility criteria, as well as the sources and methods of case ascertainment and control selection; give the rationale for the choice of cases and controls Cross-sectional study—Give the eligibility criteria, as well as the sources and methods of selection of participants (*b*) Cohort study—For matched studies, give matching criteria and number of exposed and unexposed Case-control study—For matched studies, give matching criteria and the number of controls per case
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers; give diagnostic criteria, if applicable
Data sources/measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement); describe comparability of assessment methods if there is more than 1 group
Bias	9	Describe any efforts to address potential sources of bias
Study size	10	Explain how the study size was arrived at
Quantitative variables	11	Explain how quantitative variables were handled in the analyses; if applicable, describe which groupings were chosen and why
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding (*b*) Describe any methods used to examine subgroups and interactions (*c*) Explain how missing data were addressed (*d*) Cohort study—If applicable, explain how loss to follow-up was addressed Case-control study—If applicable, explain how matching of cases and controls was addressed Cross-sectional study—If applicable, describe analytic methods taking account of sampling strategy (*e*) Describe any sensitivity analyses
Results		
Participants	13*	(*a*) Report numbers of individuals at each stage of study—e.g., numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analyzed (*b*) Give reasons for non-participation at each stage (*c*) Consider use of a flow diagram
Descriptive data	14*	(*a*) Give characteristics of study participants (e.g., demographic, clinical, and social) and information on exposures and potential confounders (*b*) Indicate number of participants with missing data for each variable of interest (*c*) Cohort study—Summarize follow-up time (e.g., average and total amount)
Outcome data	15*	Cohort study—Report numbers of outcome events or summary measures over time Case-control study—Report numbers in each exposure category or summary measures of exposure Cross-sectional study—Report numbers of outcome events or summary measures
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (e.g., 95% confidence interval); make clear which confounders were adjusted for and why they were included (*b*) Report category boundaries when continuous variables were categorized (*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period
Other analyses	17	Report other analyses done—e.g., analyses of subgroups and interactions, as well as sensitivity analyses

TABLE 7. *Continued*

	Item No.	Recommendation
Discussion		
Key results	18	Summarize key results with reference to study objectives
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision; discuss both direction and magnitude of any potential bias
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence
Generalizability	21	Discuss the generalizability (external validity) of the study results
Other information		
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based

*Give information separately for cases and controls in case-control studies and, if applicable, for exposed and unexposed groups in cohort and cross-sectional studies. Reprinted with permission.[67]

about the topic of the study and explain how their work adds to the field. Objectives should be specific. It is important when stating the objectives of a cohort study, for example, to state precisely which populations and outcomes the study evaluates.

STROBE: METHODS

The methods section should include fundamental features of the study design early on. If the study follows a cohort, that should be stated at the outset with a justification for the choice to use that design. It is especially important for a cohort study to clearly describe the characteristics of the patients and their outcomes. The creators of the STROBE checklist emphasize a particular point regarding the use of the terms "prospective" and "retrospective." They point to the ambiguity in these terms and suggest that they simply not be used. Instead, authors should offer a careful description of the ways in which data were collected and exactly when data were collected. The methods section should also include a statement of the original purpose for which the data were collected and whether that purpose differs from the purpose for which the data are being used in the particular study at hand. As was important in the abstract and introduction sections, this provides a more accurate context for the data, which is critical when judging their limitations.

Authors should be specific when describing the setting of a study. Where geographically and at what institutions were data collected? When did data collection begin and end? At what particular intervals were data collected? The methods section should clearly address these questions.

The process by which participants were selected is particularly important to a cohort study report. The eligibility criteria should be extremely clear. In a cohort study tracking the outcomes of patients who have undergone ACL revision surgery, for example, it is not enough simply to state that the patients were included if they had this procedure performed. Were pediatric patients included? Were older patients included? What if a particular patient had undergone ACL revision at an earlier date? What if the patient had rheumatoid arthritis? It is possible that the study would include all of these patients, but it must be made clear. If there are limitations on age, existing conditions, surgical indications, or any other variable, these should be stated in full. At the very least, age, gender, comorbid conditions, and diagnosis should be addressed when setting forth eligibility criteria. The authors should also report on characteristics of the broader group from which the cohort was selected. To add to the example described above, that study might have drawn all of its cohort from the population of male persons aged between 17 and 50 years who reside in Maryland. That should be included in the description of participants. Again, the goal is to provide the reader with as much information as possible that is relevant to the evaluation of the data's limitations so that the reader may judge the validity of the conclusions drawn from those data.

The authors should report follow-up methods clearly. The authors might state that questionnaires were administered at baseline and at the 6-month, 2-year, and 5-year time points. However, sometimes, questionnaires are not completed at precise time points like these. Perhaps the baseline questionnaire for one patient was in fact filled out 3 weeks after the procedure. If this is the case, the authors should justify their inclusion of the data. They might point to a study that shows that baseline question-

naires completed within one month of a procedure still provide valid data.[68]

All variables under analysis in the study should be defined unambiguously. These include outcomes, exposures, predictors, potential confounders, and effect modifiers. Disease outcomes should be identified by specific diagnostic criteria, which should also be reported. For example, it would be necessary to describe exactly what constitutes failure of an ACL reconstruction (MRI, arthroscopic evaluation, and so on). Authors should also report the source of data for each variable and how those data were measured, along with an estimate of the reliability and validity of their measurements. Ideally, measurement methods would be identical between groups being compared. If there are variations, these should be noted. Similarly, authors should address in their report any potential sources of bias. Hopefully, steps have been taken to minimize any bias present in the results. These steps should be conveyed in full to the reader.

Quantitative variables should be explained carefully, especially in terms of their grouping and the modeling of the data representing those variables. Choices are made in the process of separating continuous quantitative data representing a given variable into ordered groups, and those choices can have a significant impact on later analysis of the data. Going one step further, this applies to the choice of statistical analysis as well. Given the possibility of choosing a particular analysis to support a particular hypothesis once all of the data have been collected, that choice should be made at the outset of the study. This is true of the methods by which interactions between subgroups were examined and missing data were accounted for, as well. If these analytic methods changed over the course of the overall analysis of the data, the changes should be reported. It should also be clear in the report how confounders were selected. The Explanation and Elaboration document[67] prepared by the creators of STROBE offers a more detailed treatment of the reporting of statistical methods, which applies not only to cohort studies but to any other type of observational study as well.

Loss to follow-up deserves particular attention when reporting on a cohort study. If the total length of follow-up time is fixed, whether in terms of age or time elapsed since the baseline time point, an assumption is made during analysis: for individuals who reach this fixed endpoint without a particular outcome, there is no relation between follow-up time and the probability of their developing that outcome. Problems arise when the distribution of loss to follow-up is uneven between groups.

For example, the investigators of an ACL study hypothesize that individuals who have had an associated meniscal repair have a higher likelihood of reoperation. If 20% of the cohort is lost to follow-up, including 80% of the meniscal repair patients, it is difficult to establish a potential relationship between the meniscal repair and the outcome, a reoperation. It may appear as though there is no relation when further observation of those individuals may have established that very relation. For this reason, members of the cohort who were lost to follow-up must be distinguished from those who remain under observation until the fixed endpoint of the study. Those lost to follow-up could be excluded from the study or they could be treated as though they withdrew without the outcome in question, either at the end of the study or on the date on which they were actually lost to follow-up. When planning the study, investigators should determine how loss to follow-up will be handled. However they choose to treat those lost to follow-up when analyzing their data, they should report the exact number falling into this category.

STROBE: RESULTS

Another item on the STROBE checklist that concerns cohort studies in particular hinges on the reporting of the timing of follow-up. This should be addressed in the results section. These data can be summarized as a mean or median duration of follow-up in combination with a total of the person-years observed. For example, ACL surgery investigators might report that follow-up lasted a mean of 5.2 years in 31 patients, for a total of 161.2 person-years of follow-up. Minimum and maximum follow-up times can be used in conjunction with percentile distribution for a more comprehensive picture of follow-up duration.

Closely related is the STROBE checklist's suggestion that outcome rates should be presented by follow-up time point, each time point being associated with a certain number of outcomes and a certain number of person-years of follow-up, such that the rate can be clearly shown as a ratio between the number with a certain outcome and the total number of person-years observed at that follow-up time point. For example, the investigators might report that 6 failures had occurred by the 1-year time point, at which point 31 person-years of follow-up had been performed, for a rate of about 2 failures per 10 person-years. In addition to mean data on follow-up times,

investigators should report just how many participants were involved over time. A flowchart can be quite useful in illustrating the answers to the following questions at various points in the study's progress: Who might be eligible to participate? Who has been confirmed as eligible by examination? Who is actually participating in the study? For those eligible who are not participating, what is the reason? Who is in the process of follow-up? Whose information has already been analyzed?

The characteristics of those participating in the study must be provided in the results section just as they were in the methods section, but in tabular form. This allows the reader to judge the study's generalizability. In cohort studies, exposures and potential confounders should be reported by group. If participants are missing data for certain variables, the numbers missing should be listed by variable.

When one is reporting the main results, estimates should be given both unadjusted and adjusted for confounders. The latter estimates require a figure denoting their precision, such as a confidence interval. As mentioned earlier when explaining the reporting of quantitative variables in the methods section, any continuous variables that are converted to ordinal variables should be reported with the boundaries of those categories. Of course, all analyses should be reported, including those of subgroups and interactions.

STROBE: DISCUSSION

When composing the discussion section, authors must be careful to separate their own opinions from their rigorous and unbiased interpretation of the data. The former have no place in the discussion. In keeping with a recurring theme, the discussion section should address the limitations of the study while summarizing the most important results in terms of the original objectives of the study. In the methods section, the authors should have described the measures they took to guard against the effects of potential biases. Were these measures successful? If bias appears to have affected the results, to what extent and in what direction did this happen? These are questions that should be answered in the discussion.

All things considered—biases, statistical uncertainty, the very nature of the study—what does the study show? Every other element of the discussion section serves to help answer this question. It cannot be emphasized strongly enough that this interpretive task is easily clouded by authors' personal opinions. To successfully craft the discussion section, authors must devote attention to this tendency, such that its effects might be reduced.

Finally, the authors must broaden the scope of the discussion to explain the generalizability of their results. At this point, they have offered an interpretation of the study's findings within the realm of the study itself, but how do their findings apply to patients outside of the study cohort? It is in this part of the discussion that the study's impact on clinical practice can become clear. Previous explanation of the setting of the study, eligibility criteria for participation in the cohort, and the exposures and outcomes to be measured help readers to assess on their own the generalizability of the study findings. Naturally, authors should address these topics when offering their own argument for the ways in which the findings can be applied in other circumstances. Also important to this critical process is the disclosure of 2 more factors that may introduce bias: sources of funding for the study and any conflicts of interest the investigators may have.

THE PURPOSE OF STROBE

The reader of a cohort study report should be able to critically evaluate that report in 2 ways: in terms of its internal validity and in terms of its external significance. Do the data clearly support the conclusions reached regarding the specific domain of the study? Do those specific conclusions support the broader implications the authors suggest? After all, the ultimate goal of most cohort studies is to provide information that will make a positive impact on clinical practice.

Authors who adhere to the STROBE checklist ensure that their readers have the tools necessary to make these critical judgments. Each element of the checklist serves this purpose, some more obviously than others. Attention to this guiding principle should help authors effectively execute the particular items presented here.

While offering an overview of the whole STROBE checklist, this chapter has focused on those items with unique application to cohort study reporting, particularly in orthopaedic surgery. For a comprehensive discussion of the application of the STROBE checklist, investigators may consult the STROBE Explanation and Elaboration document referenced above.[67]

Brian W. Boyle, B.A.
Michael Soudry, M.D.
Robert G. Marx, M.D., M.Sc., F.R.C.S.C.

SECTION 6

Study Designs: Case-Control Studies, Level III Evidence, STROBE Checklist

Clinical investigations are an integral component to assessing and improving the care of our patients. Whereas prospective studies are considered the "gold standard" of clinical outcomes research (type I and type II studies), it has been estimated that 60% of surgical questions cannot be answered by these methods. Case-control studies (type III) are a type of observational study that can be quite useful in identifying risk factors associated with a specific outcome. In this type of study, subjects with the outcome of interest (cases) are compared with similar subjects who do not have that outcome (controls). Data from each subject's treatment records are then compared to identify factors common to the cases but not common to the controls by use of epidemiologic and statistical methods. These factors may be genetic, gender, or chemical or based on exposure or other comorbidities. Case-control studies are most useful when the research question addresses outcomes that are rare or take a long time to develop.[69] In these situations randomization or prospective cohort studies may not be feasible because of the required length of follow-up and expenses. Case-control studies are also indicated when randomization may be unethical (such as when investigating the fracture risk associated with the use of nonsteroidal anti-inflammatory drugs[70]).

The classic case-control study was performed by a young medical student, Ernst Wydner, who was fascinated with lung cancer. He interviewed a pool of 649 lung cancer patients (cases) and 600 patients with other cancers (controls) and found that the incidence of lung cancer was 40 times higher in smokers that in those who did not smoke.[71] Other investigators who read this index study began studies of their own to further understand the association between smoking and the development of lung cancer.

Case-control studies have been very useful in orthopaedics and sports medicine to assess the risk of a specific injury or to assess the risk of a certain outcome after injury or surgery. The results of these studies can lead to strategies to reduce the risk of injury or to improve the clinical outcomes of our treatments.

The focus of this chapter is to provide the investigator a structure to assist him or her in designing and carrying out a case-control study.

WHEN TO CONSIDER A CASE-CONTROL STUDY

When considering a case-control project, the investigator should consider several critical points. The advantages and disadvantages of this type of study should be assessed before beginning the study to be sure that the outcome will answer the research question. Table 8 provides some of the advantages and disadvantages of a case-control study.

First, the topic of study should be one that is familiar to the research team. Research questions (see be-

TABLE 8. *Strengths and Limitations of a Case-Controlled Study*

Strengths
Facilitates the study of rare outcomes
Facilitates the study of conditions with substantial time between exposure and outcome
Control groups can be matched according to known (or suspected) confounding variables
Allows for the study of multiple potential causes of an outcome of interest
Relatively inexpensive
Can be completed over relatively short time periods
Limitations
Inefficient when the exposure is rare
Information on exposure and history that is derived from interview is subject to recall bias
Selection of an appropriate control group may be challenging
Lack of randomization means that groups may suffer from an imbalance of confounding factors
Can only study one outcome of interest
Validation of exposure information is often difficult (or impossible)
Cannot provide information on prevalence of the outcome of interest
Unable to establish causality
Methodology and correct interpretation of results may be challenging

Reprinted with permission from Busse JW, Obremskey WT. Principles of designing an orthopaedic case-control study. *J Bone Joint Surg Am* 2009;91:15-20 (Suppl 3).

low) are best formulated by an investigator who has a thorough understanding of the research topic and desires to study one specific unanswered question. The number of cases available for inclusion should be assessed and be adequate. Studies that are based on relatively few cases may be useful for evaluating rare conditions but may not lead to firm conclusions. Case-control studies are also retrospective in nature and rely heavily on previously collected data. Thus, when choosing an area of interest, it is critical that a database is easily accessible. Finally, once the study design is complete, the research team should perform an assessment of its ability to complete the study, the amount of time it will take, and whether the desired outcome will be achieved.

FORMULATING A RESEARCH QUESTION: THE PICOT FORMAT

The first step in conducting a research study is to pose a study question, and it is arguably the most important step. Spending adequate resources to develop a clear and relevant question will "determine the research architecture, strategy, and methodology."[72] The research question should be framed in a manner that is easily understood. A poorly designed question can hinder your research efforts, making it difficult for readers to interpret the results, and ultimately, jeopardize publication.

One way to enhance the organization and clarity of the research question is to use the PICOT format.[73] When using the PICOT format, one frames the study question in terms of the population of interest, the intervention, the comparator intervention, outcomes, and the time frame over which the outcomes are assessed.[73] In case-control studies, the population should be specific, addressing key eligibility criteria such as type of patient, geographic location, and qualifying disease or condition. The intervention is one or more exposure variables under investigation, and the comparator is often the absence of those factors. The outcome is the proportion of cases exposed to the variables under question compared with the controls. The data collected are usually reported as odds ratios. It is worth mentioning that the PICOT format is generally most useful for comparative studies or studies of association between exposure and outcome.[73]

Consider the following example: a researcher wants to investigate whether a traumatic anterior shoulder dislocation can increase the risk of severe shoulder osteoarthritis (OA) developing in later life. Marx et al.[74] designed an elegant, case-control study to evaluate this question. Their cases (n = 80) comprised patients who had had either a hemiarthroplasty or total shoulder arthroplasty for OA of the shoulder. They chose this group in that each had severe OA requiring replacement surgery, the diagnosis was easily confirmed at the time of surgery, and the sample of patients was easily identifiable. They excluded patients with rheumatoid disease, avascular necrosis, cuff tear arthropathy, and other systemic causes of severe shoulder pain.

Marx et al.[74] chose a group of patients undergoing total knee replacement (n = 280) for OA of the knee without OA of the shoulder as the control group because this group of patients had similar age, gender, and comorbidity distributions and was also easily identifiable. Subjects were then asked whether they had ever had a shoulder dislocation. The findings of this study were that the risk of shoulder OA developing was 19.3 times greater if there had been a shoulder dislocation in earlier life. The reader is encouraged to read this study as an elegant example of a clinical case-control study.

IDENTIFYING POTENTIAL RISK FACTORS

In general, the investigator has already identified potential risk factors, or exposures, that may have an association with the outcome when considering the study design. In the first example, Marx et al.[74] drew upon their clinical experience when asking the question about a possible association between shoulder dislocations and later development of shoulder arthritis. In other cases there may be one or more of a list of potential risk factors that may have an association with the identified outcome. In either scenario it is vital that a reasonably complete database is identified that can be easily searched for potential risk factors. It does little good to formulate a research question only to find that the necessary data are either difficult to obtain, incomplete, or simply not available.

IDENTIFYING THE CASES

When designing a case-control study, investigators begin by selecting patients with the outcome of interest, the case patients. The enrollment criteria for the case patients must be well-defined and as specific as possible. Criteria may include age, gender, and/or geographic location. The investigators must specify how the presence or absence of the desired outcome to be studied is established (e.g., clinical symptoms,

physical examination findings, imaging studies, or laboratory studies).[69,75] It is preferable also to define the time period of collection because diagnostic criteria can change over time. Detailed descriptions of the case participants will aid in determining the validity of the study results.[75] For example, in a study looking at the fracture risk associated with nonsteroidal anti-inflammatory drugs, acetylsalicylic acid, and acetaminophen, the investigators identified fracture cases through the National Hospital Discharge Register in Denmark between January 1, 2000, and December 31, 2000.[70]

IDENTIFYING APPROPRIATE CONTROLS

The next step is to identify the controls—that is, the group of individuals who are reasonably similar to the cases but in whom the outcome of interest has not occurred. The controls are usually selected from the same population as the cases so that the only difference between the 2 groups is the exposure to the putative risk factors.[69,75] Similar to case assessment, the method of control selection should be clearly documented.[75] For example, in a study examining the risk of ACL tearing based on ACL volume, tibial plateau slope, and intercondylar notch dimensions as seen on MRI, investigators compared the MRI findings of 27 patients who had had a noncontact ACL injury with controls who had an intact ACL and were matched by age, gender, height, and weight.[76]

Sometimes, the rarity of the disease under investigation may limit the total number of cases identified; in these situations, statistical confidence can be increased by selecting more than 1 control per case.[69] Typically, the ratio of controls to cases should not exceed more than 4:1 or 5:1.[69]

DATA COLLECTION

Once the appropriate cases and controls have been selected, the investigators look back in time to examine the relative frequency of exposure to variables in both groups. The collection of data may involve a chart review or patient interviews. Whenever possible, data collection should be done by study personnel who are blinded to patient status—that is, whether the patient is a case or control.[77] This will limit the possibility that the information is collected differently based on patient status.[77] These data will then allow the calculation of a measure of association between the exposure variables and the outcome of interest. A flow diagram of how the data should be collected is shown in Fig 4.

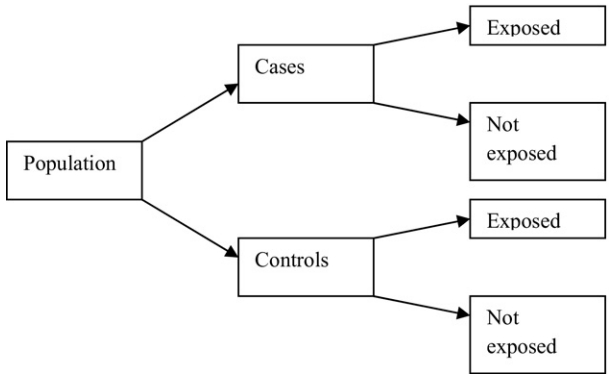

FIGURE 4. Diagrammatic representation of a case-control study. Investigators begin by identifying cases and suitable controls. In a retrospective manner, the cases are compared with controls for the presence of risk factors or past exposures.

It is important to keep track of your study's raw data throughout its progress to ensure accuracy and to strengthen the reporting of your case-control study.[73] A database of this information helps facilitate the process. The success of your study also depends on a qualified and experienced research team, because you will simply not have enough time to complete the project on your own.[69] One particularly important study personnel to include on your team is a research coordinator. This person is responsible for organizing the trial and communicating with the principal investigator, providing details on patient recruitment, data submission, and any problems experienced.[69]

INSTITUTIONAL REVIEW BOARD

Most case-control studies involve the collection of personal patient data. As a result, an approval by the institutional review board and ethics committee will likely be required before beginning your study. In general, the application usually involves specifying the details of your research, including the question, methodology, statistical analyses, and outcomes of interest. It will also request a copy of the informed consent form that will be read and signed by patients before study participation. Finally, it may ask for a description of the estimated study budget.

STATISTICAL CONSIDERATIONS

Power Analysis

Sample size is an important consideration when designing your case-control study. An appropriate

Table 9. *Basic 2 × 2 Table Illustrating How to Calculate an Odds Ratio*

	Disease	
	Yes	No
Exposure		
Yes	a	b
No	c	d

NOTE. The odds ratio is given by (a/c) ÷ (b/d) or ad/bc.

sample size ensures that your study is "powered" to detect a difference when there is one.[69] Details about the sample size calculation should be reported in the final publication as well.[75] Although information on how to calculate sample size is beyond the scope of this chapter, investigators are advised to consult epidemiologic or statistical textbooks for further details. This leads to another critical consideration when conducting a case-control study: the use of biostatisticians, who will be responsible for the appropriate statistical analyses. If necessary, involving a biostatistician early on in the planning phases of your study may be helpful.

Data Analysis

In case-control studies, a measure of association between the exposure(s) and the target outcome is usually reported as an odds ratio.[69] This refers to the odds of an event occurring in the exposed group compared with the odds of the same event happening in the unexposed group.[69] The final value can range from 0 to infinity. One is the neutral value, which means that there is no difference between the 2 groups.

Table 9 illustrates a basic 2 × 2 table. The odds ratio is given by (a/c) ÷ (b/d) or ad/bc.

For example, in a case-control study looking at the risk factors for plantar fasciitis, the investigators found an odds ratio of 3.6 for those who reported that they spent the majority of the day on their feet.[78] This indicates that the odds of weight bearing for most of the day is 3.6 times higher in patients diagnosed with plantar fasciitis than in those who do not have the disease.

LIMITING BIAS IN THE CONDUCT OF A CASE-CONTROL STUDY

As alluded to previously, case-control studies are retrospective in nature and often rely on patients' recollections to identify exposure, making them susceptible to recall bias.[69] This occurs when patients with an adverse outcome have a different likelihood of recalling past exposures than those who have not had an adverse outcome.[69] It is often difficult to limit recall bias in case-control studies. One way is to study past exposures that are objective and can be easily confirmed. If exposure data are being collected by study personnel through patient interviews, the assessors should also be blinded to the status of the patient (i.e., whether the patient is a case or a control) so that the information is not collected differently.[77] For example, Marx et al.[74] used glenohumeral dislocation as the exposure variable. They believed that it would be very unlikely for patients to incorrectly recall whether they had ever had a shoulder dislocation in the past. Furthermore, they attempted to confirm the exposure data by contacting patients and eliciting additional information such as date of dislocation, mechanism of injury, and number of recurrences.

Another important source of bias is from confounders—that is, a variable that is associated with both the exposure and the outcome. In case-control studies, the control group is selected so that it is ideally similar to the cases, except for the exposure status. However, any control group is at risk for an unequal distribution of prognostic factors compared with the cases, which can lead to biased results.[77] Careful selection of appropriate control patients is an important way to limit the effects of confounding variables. In the study by Marx et al.,[74] for example, the authors chose a group of patients who had undergone total knee arthroplasty because they were similar to the cases with respect to age, health, and mental status. They also identified prior surgery for recurrent shoulder dislocation as a potential confounding variable in the study. As a result, they conducted a subgroup analysis by excluding patients with prior surgeries, which is another way to strengthen the reporting of the case-control study.[75]

REPORTING A CASE-CONTROL STUDY

When preparing the manuscript for publication, it is important to maintain adequate transparency of your study.[75] The reporting of your case-control study should be detailed enough to allow readers to assess its strengths and weaknesses.[75] Investigators are strongly encouraged to refer to the STROBE statement[75] for further details on the reporting of observational studies to improve the overall quality of the final manuscript. Table 7 in section 5 (pp 24-25) also provides a checklist of items from the STROBE statement to include in the publication.

In the manuscript, the "Introduction" section needs to address the reasons for the study and the specific

objectives and hypotheses.[75] Next, the "Methods" section should provide details on the study's processes. The goal is to provide sufficient information so that the readers can judge whether the methods were able to provide reliable and valid answers.[75] In case-control studies, it is important to document the eligibility criteria for study participants, including the method of case ascertainment and control selection.[75] All study outcomes should be clearly specified as well, including the diagnostic criteria.[75] Furthermore, you should describe the statistical methods used in the study and how the sample size was calculated.[75]

The "Results" section is a factual account of the study's outcomes, which means that it should not reflect the author's views and interpretations.[75] Data should be provided on the recruitment and description of study participants. It is also important to explain why patients may not have participated in the study or why they were excluded if applicable; this allows the readers to judge whether bias was introduced into the study.[75] The main outcomes should be documented, including the numbers in each exposure category and the statistical analyses.

In the final stages of the manuscript, the "Discussion" section addresses the issues of validity and meaning of the study.[75] A structured approach has been suggested by the STROBE statement, which involves presenting the information in the following manner: (1) summarize key findings, (2) provide possible explanations or mechanisms, (3) compare current outcomes with the results from previous studies, (4) list study limitations, and (5) specify the clinical and/or research implications of current study findings.[75]

CONCLUSIONS

In the hierarchy of evidence, case-control studies represent Level III evidence.[69] However, despite some methodologic limitations associated with case-control studies, they can be very useful in informing many research questions, particularly when they are well-designed and -reported.

Kevin Chan, M.D.
Kevin P. Shea, M.D.
Mohit Bhandari, M.D., Ph.D., F.R.C.S.C.

SECTION 7

Study Designs: Case Series, Level IV Evidence

Although RCTs provide the highest level of evidence, they are also the most expensive studies to conduct.[79] As patients become more educated, it is also more difficult to enroll patients because they are looking for specific treatments and are less willing to risk being in a control group. Even in cases where it is not clear what the best treatment is for a given patient population, patients want to be increasingly involved in an informed decision-making process and may opt for a more aggressive treatment strategy to maximize the possibility of improving function.

We cannot ignore the ethical question posed by many surgeons when considering randomized trials.[80] Is it ethical for a surgeon to offer a patient no treatment when no surgical treatment is considered inferior treatment by the surgeon? If the surgeon is uncertain whether one treatment is better than the other (clinical equipoise),[80] then patients can be enrolled. If the surgeon is certain or not completely uncertain that the treatment to be studied is superior, then ethically, the surgeon should not perform the inferior treatment (control treatment) on a patient. A case series may be the preferred type of study in this instance.

The difficulties with RCTs have resulted in more case series being performed.[80-82] In a recent statement, the editors of *Arthroscopy* acknowledged that case series are the most common type of article in their journal.[80] However, they pointed out that not all case series are alike, and when properly performed, case series can improve patient care and add to our clinical knowledge.

There are 2 common types of case series designs. These are studies with prospectively collected data with retrospective data analysis (Fig 5A) and retrospective reviews of cases (Fig 5B). Before initiating any study involving patients, approval from the institutional review board or ethics board is required.[83] For investigators in

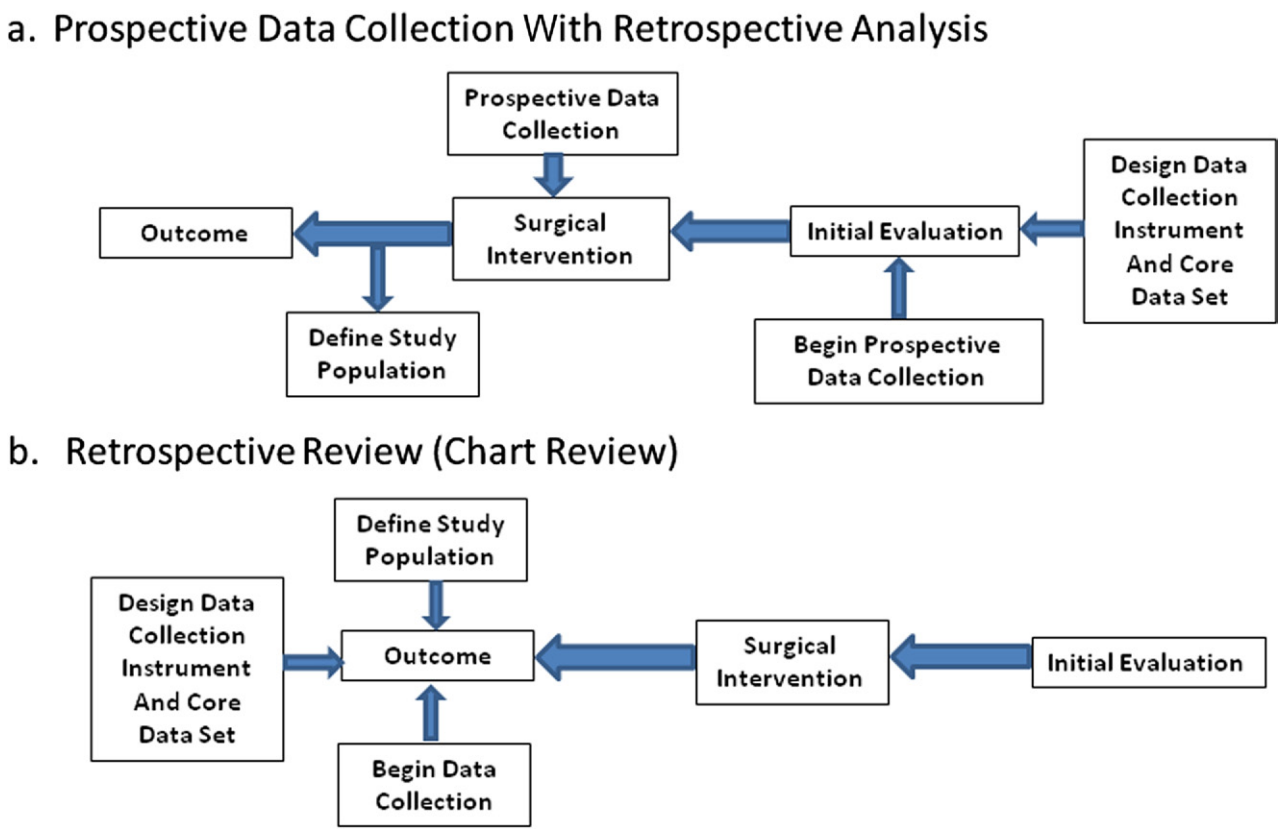

FIGURE 5. Prospective data collection versus retrospective chart review. In a study with retrospective analysis of prospectively collected data, data collection is started preoperatively. For retrospective chart review, data are collected postoperatively.

the United States, it is also important to address Health Insurance Portability and Accountability Act privacy guidelines before initiating any study.[84] This chapter will describe each study and provide suggestions of how to improve the quality of the study. With proper planning and study design, case series can provide an important addition to the orthopaedic literature.

RETROSPECTIVE REVIEW OF CASES

A retrospective review involves gathering data that have been collected for reasons other than research. These are most commonly seen in the literature describing uncommon pathologies or procedures. By reviewing past cases, these procedures can be added together over several years and studied.

Most retrospective review studies involve review of patient charts.[82] With the advent of electronic medical records, these studies may be easier to perform and more complete than in the past. The main weakness of these studies is that there was no consistent data collection plan for all patients before the study was initiated. The quality of data is dependent on the patients' charts and dictations, which may have changed over the years of the studies. Many studies use retrospective data to determine the patient population and then prospectively collect outcome data on these patients. These studies cannot show improvements over time, but they can give a general overview of the outcome of specific procedures.

The first step in designing a retrospective review is to establish the patient population. After the specific procedure that the study will be based on has been chosen, patients are identified. Most chart reviews are done by searching billing software for specific procedure codes. This provides an overall list of patients; however, all procedures should be verified by the operative notes. It is also important to have 2 sources to search for patients. For example, studies have used billing software and the physician's personal log.[85] This is especially important if there are multiple physicians and the study covers multiple years. In addition, it should not be assumed that all physicians code their procedures the same.

Table 10. *Keys to a Quality Retrospective Chart Review*

1. Determine how study patients will be identified. It is important to identify all patients with intervention that is being studied. If only a small subset of patients is used, the data may not represent the actual outcome.
2. Define in detail the inclusion/exclusion criteria and strictly enforce them. No patient should be removed from the study unless he or she has a specific exclusion criterion.
3. Have strict guidelines for data collection. Do not allow a data point to be assumed negative just because it is not in the chart.
4. Define what is an acceptable level of complete data, as well as which data points are mandatory, before data collection.

Strict inclusion and exclusion criteria will improve the quality of the study. If patients aged under 18 years are included in the study, additional consents may be required from one's ethics committee. Gender should rarely be used as an inclusion/exclusion criterion in orthopaedic studies. A specific time frame for which patients will be included must be established. This time frame should be based on when the procedure was performed in a similar manner over time. In addition, if changes in postoperative or rehabilitation protocols were noted, this should either be noted as a data point or accounted for by the time period selected.

After the inclusion/exclusion criteria have been determined, it must be determined what data points will be collected from the charts. These data points should be points that are absolute and do not need to be interpreted by the chart reviewer (Table 10). These data points should also be points that can be expected to be found in the majority of patient charts. For instance, if knee pain is a data point, then the most consistent way to gather these data would be a yes/no selection. Many charts may list it on a scale of 1 to 10, but some may list it as mild to extreme. If data are not presented in the same format, it is better to dichotomize the data rather than trying to make 2 scales fit. In addition, reviewers should not assume values for data. For example, if pain is not mentioned, it cannot be assumed that pain is absent. This would have to be left as a blank data point. For data that may be kept by other departments (i.e., radiology), the availability of the data needs to be determined. If radiographs are necessary for the study, they must be available for all patients in the study. In some institutions radiographs may only be kept for 10 years, and old radiographs may be destroyed or archived in difficult-to-access sites.

It is common in a retrospective review for one person to collect all the data from the charts. This means the data are collected consistently; however, it may also add a single reader's bias. For data collection, we suggest that a specific data collection form be designed before initiation of data collection to reduce any bias. This form will define data points and direct data collection. This will also reduce the need to interpret nonspecific data. Before starting data collection, the investigators must decide what percentage of data points are needed to include patients in the study. For example, if the data sheet has 20 data points to collect and 80% is the level of data that must be collected to be included in the study, then any patient who is missing more than 4 data points would not be included. However, some data points should be mandatory, especially if they involve the question the study aims to answer.

STUDIES WITH PROSPECTIVELY COLLECTED AND RETROSPECTIVELY ANALYZED DATA

To track longitudinal patient outcomes data, patients complete questionnaires before intervention and then at specific time points. These data are commonly stored in a research database and are considered prospectively collected. These data are collected on consecutive patients with a predetermined survey instrument that is completed by all patients. These studies suggest clinical course and response to intervention.

Data collection instruments are developed to cover all procedures done on a specific joint. For example, we have a knee arthroscopy outcome instrument (Figs 6-8), a shoulder instrument, and a hip arthroscopy instrument. All patients who are seen in the clinic complete one of these instruments. In addition, physical examination findings, surgical findings, and treatments are recorded. At defined time points, follow-up questionnaires are collected from patients. Because the study question is not designed at the beginning of data collection, the available data for the study will be limited to the specific instruments that were implemented and collected prospectively. Before starting prospective data collection, the data instruments should be carefully developed. You should develop these based on what you think will be important in 2 years, 5 years, 10 years, and beyond. Leaving off one key data element can limit the productivity of your research database. A thorough review of the literature will also help determine which data points are important.

The data instrument should be a comprehensive assessment of outcomes after a treatment, which includes a generic measure of health-related quality of life, a condition-specific measure of function, and a measure of patient satisfaction with outcome.[86,87] Any scoring sys-

FIGURE 6. Example of questions from a physician-completed knee objective data collection form.

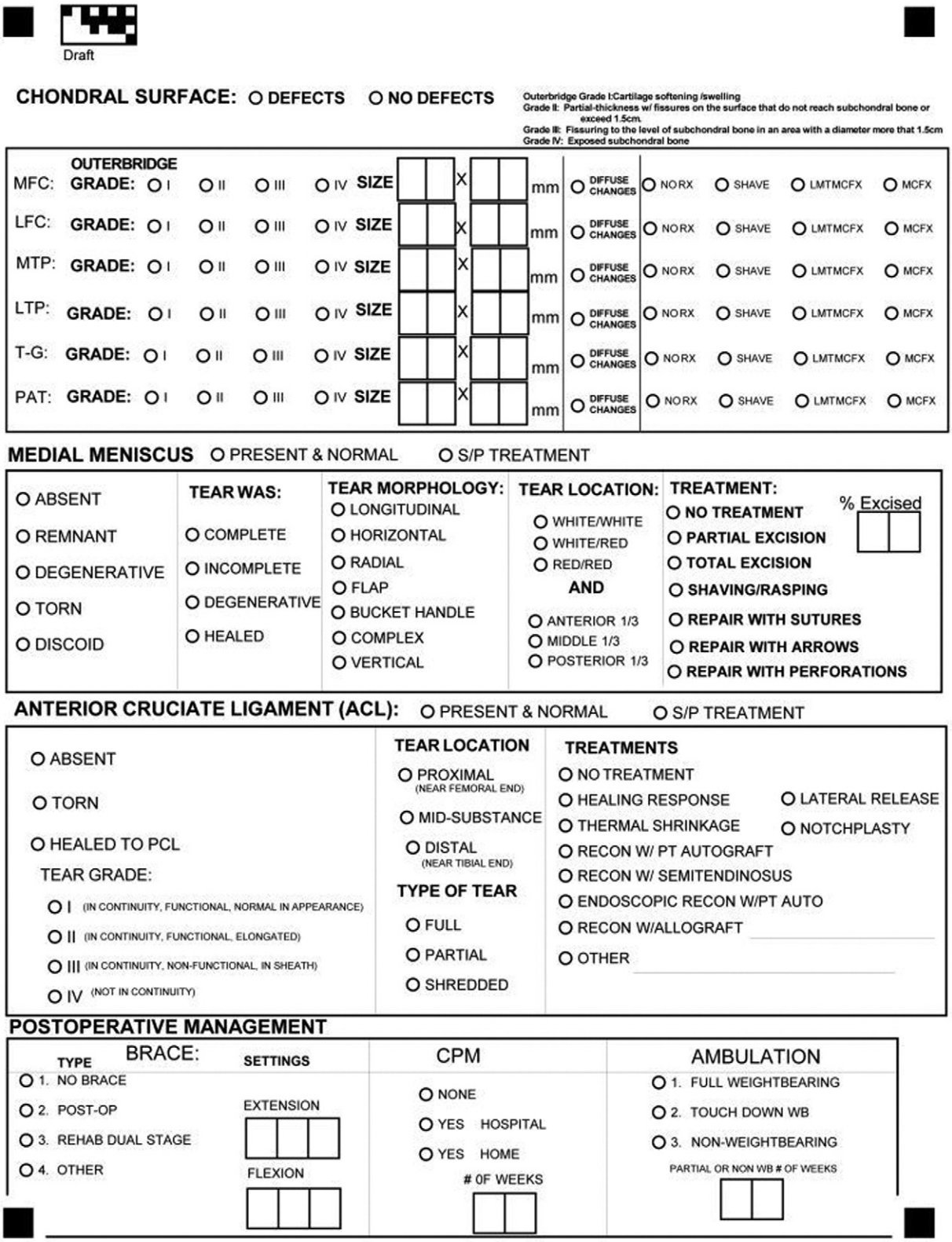

FIGURE 7. Example of questions from a physician-completed knee surgery data collection form.

Please grade each symptom that you experience currently during your highest level of activity

Swelling:	O None O Mild (on severe exertion) O Moderate (on ordinary exertion) O Severe (constant)	
Pain:	O None O Inconstant and slight during severe exertion O Marked during severe exertion	O Marked on or after walking more than 2 km O Marked on or after walking less than 2 km O Constant
Crutch Use:	O None O 1 Crutch (stick or crutch) O 2 Crutch (stick or crutch) O Weight bearing impossible	
Walk with Limp:	O Yes (severe or constant) O No (none) O Somewhat (slight or periodical) O Not Applicable	
Locking:	O No locking and no catching sensations O Catching sensations but no locking	O Locking frequently O Locking occasionally O Locked joint
Instability:	O Never giving way O Rarely during athletics or other severe exertion O Frequently during athletics or other severve exertion	O Occasionally in daily activities O Often in daily activities O Every step
StairClimbing:	O No problems O Slightly impaired O One step at a time O Impossible	
Squatting:	O No problems O Slightly impaired O Not beyond 90 degrees O Impossible	

14. Currently, are you back to your original fitness program? O Yes O No O Somewhat O Not Applicable

15. Please choose one from the following which best describes your current activity level.

O Level 10	Competitive Sports(Soccer, Football, Rugby (national elite)
O Level 9	Competitive Sports(Soccer, Football, Rugby (lower divisions), hockey, wrestling, gymnastics)
O Level 8	Competitive Sports(Racquetball, Squash, Track and Field, Alpine Skiing)
O Level 7	Competitive Sports (Tennis, Athletics(Running), Handball, Basketball, Motorcross, Cross county track Recreational Sports (Soccer, Football, Hockey, Squash, Athletics(jumping), Cross country track)
O Level 6	Recreational Sports (Tennis, Handball, Basketball, Alpine skiing, Jogging 5X/week)
O Level 5	Work (Heavy Labor) Competitive Sports (Cycling, X-country Skiing) Recreational (Jogging on uneven ground 2x/week)
O Level 4	Work (Moderately Heavy Labor (truck driving, etc) Recreational Sports (Cycling, Cross Country Skiing, Jogging on even ground 2X/week)
O Level 3	Work (Light Labor) Comp & Rec Sports (Swimming), Hiking, Backpacking
O Level 2	Work (Light Labor) Walking on uneven ground possible but impossible to backpack or hike
O Level 1	Work(light labor) Walking on even ground possible
O Level 0	Sick leave or disability pension because of knee problems

17. Have you had any further surgery on your affected knee since your latest surgery ***that was performed elsewhere?*** O Yes O No

If so, when? DATE: ☐☐ / ☐☐ / ☐☐☐☐

If so, what Procedure?

O Ligament O Arthroscopy/Debridement
O Meniscus O Cartilage O Other

Satisfaction

18. Rate the following on a scale from 10 to 1.

	Very Satisfied									**Very Unsatisfied**
2) How satisfied are you with your current **OUTCOME** from your knee surgery?	O 10	O 9	O 8	O 7	O 6	O 5	O 4	O 3	O 2	O 1

FIGURE 8. Example of questions from a patient-completed knee subjective data collection form.

Table 11. *Keys to Prospectively Collected Data and Retrospective Analysis Studies*

1. When building your data collection instruments for prospective data collection, use validated outcome scores including a condition-specific score, a quality-of-life scale, and a measure of patient satisfaction.
2. Maintain your database. Set rules and protocols to ensure quality data are maintained.
3. We recommend that patient follow-up forms be mailed on an annual basis. You may only need 2-year and 10-year follow-up, but annual data collection allows you to keep in touch with patients and also provide data on improvement or decline over time.
4. Identify the study group from the database using inclusion and exclusion criteria. If patients refuse to participate, this must be respected.

tem that will be used as part of the data instruments must have been tested to determine whether it can measure change after an intervention. Because the instruments are picked before data collection, it is very important that a valid, reliable, and responsive score is used to collect data (see section 9). If an untested score is used, when the data are analyzed and the results are poor, one will not know whether patients did not improve or whether the instrument was not able to detect the improvement over error. Valid and reliable questionnaires will ensure quality data collection.

If a database is set up for the collection of data over a series of years, steps must be taken to ensure that the data are of the highest quality (Table 11). Standard operating procedures for data collection, data entry, and data verification must be developed and implemented. In addition, Health Insurance Portability and Accountability Act guidelines must be followed in the United States when collecting and storing data.[84] Data audits should also be performed annually or every other year. For a database to be useful, it must be filled with accurate, quality data.

Based on what data have been collected, a study can be designed. Just as in a retrospective study, the inclusion/exclusion criteria are crucial in these studies. However, in studies where all data are collected prospectively, incomplete data should not be exclusion criteria. In addition, with large numbers, more variables can be studied. With regression analysis of large groups, independent predictors of the outcome can be determined.

Once the inclusion/exclusion criteria have been determined and the variables of interest are determined, the database can be queried to extract the data. A control group may also be identified by use of the same inclusion/exclusion criteria but with a previously performed technique or with surgery that did not include the procedure of interest. Care must be taken when identifying a control group. It must be of equal trauma to the patient, equal recovery, and equal rehabilitation. If the control group does not quite match up, it is better to proceed without a control group.

When data have been queried and put into a spreadsheet, continuous data should be analyzed for normal distribution. After this, data can be analyzed by use of the proper statistical tests.[88] These tests will be discussed in further chapters. We encourage all researchers who are starting to perform clinical studies to obtain input from a statistician. It is also helpful to have an independent statistician review all data analysis at the completion of the study.

PRESENTATION OF DATA

When presenting data from these studies, it is crucial to fully describe how data were collected. There are many examples in the literature of studies that are described as retrospective reviews but it is unclear how the patients were identified and how additional data were obtained. Readers are more likely to consider your study if it is easy to understand the study design. It is also very important to include the numbers of patients in the study. This should start with the total number of patients who had the procedure completed. Then, the number of patients who fit the inclusion/exclusion criteria should be listed. Failures should also be reported. Failures should be adequately defined. If patients are followed up to an endpoint, this should also be defined. If it is unclear why patients are considered failures, then it will be difficult for readers to understand the study outcome. Regarding follow-up, the number of patients available for follow-up should be reported. Then, the percentage of those patients in whom follow-up was obtained should be reported. In some studies, having follow-up is considered an inclusion criterion. If the readers do not know whether these data are on 80% of the patients or on 30% of the patients, then it is again difficult for them to interpret the study outcome.

CONCLUSIONS

Case series are common in the literature today (Table 12). Many of these studies use prospectively collected data, which increases the quality of these studies. As more physicians begin to monitor their patients

TABLE 12. *Summary*

1. Case series, when done properly, are important additions to the literature.
2. Prospective data collection allows for quality research with minimum selection bias. It also allow physicians to track all of their patients over time. This provides a means of patient feedback and improving patient care.

for quality-of-care purposes, more of these prospective database studies will be completed. If these studies are well-designed and well-executed and the analysis is done properly, then they provide important information to the literature. Depending on the individual clinical setting, this type of study could become the research study of choice.

Karen K. Briggs, M.P.H.
Robert F. LaPrade, M.D., Ph.D.

SECTION 8

Special Designs: Systematic Reviews and Meta-Analyses

Health care professionals are increasingly required to base their practice on the best available evidence derived from research studies. However, these studies may vary in quality and produce conflicting results. It is therefore essential that health care decisions are not based solely on 1 or 2 studies but rather take into account the range of information available on that topic.[89] Health care professionals have traditionally used review articles as a source of surmised evidence on a particular topic because of the explosion of medical literature and scarcity of time.

Review articles in the medical literature are traditionally presented as "narrative reviews," in which experts in a particular field provide a summary of evidence. There are several key disadvantages to the use of traditional narrative reviews. The validity of a review article is dependent on its methodologic quality.[89] Authors of narrative reviews often use informal, subjective methods to collect and interpret studies and are therefore prone to bias and error.[90] Reviewers can disagree on issues such as what types of studies to include and how to balance the quantitative evidence they provide. Selective inclusion of studies to reinforce preconceived ideas or promote the author's view on a topic also occurs.[89,90] Furthermore, traditional reviews often ignore sample size, effect size, and research design and are rarely explicit about how studies are selected, assessed, and analyzed.[90] In doing so, they do not allow readers to assess the presence of potential bias in the review process.[90]

In contrast to narrative reviews, systematic reviews apply "scientific strategies in ways that limit bias to the assembly, a critical appraisal, and synthesis of relevant studies that address a specific clinical question."[89]

WHAT IS A SYSTEMATIC REVIEW?

Systematic reviews are scientific investigations conducted with a specific methodology using independent studies as "subjects."[91] They synthesize the results of multiple primary investigations using established strategies aimed at limiting random error and bias.[91] Strategies include a comprehensive search of relevant articles using explicitly defined and reproducible criteria. In a systematic review, primary research designs and study characteristics are appraised, data are synthesized, and results are interpreted.[91]

Systematic reviews can be quantitative or qualitative in nature. In a qualitative systematic review, results of primary studies are summarized without being statistically combined. Quantitative reviews, on the other hand, are known as meta-analyses, which use statistical methods to combine the results of 2 or more studies.[91]

Current evidence-based practice guidelines are based on systematic reviews appropriately adapted to local circumstances and values. Economic evaluations compare the costs and consequences of different courses of action. The knowledge of consequences available for these comparisons is often generated by systematic reviews of primary studies.[91] In this manner, systematic reviews play a key role in clinical decision making by allowing for an objective appraisal of knowledge accumulated from the robust and

Table 13. *Features of a Systematic Review*

Key Points
Systematic reviews address a specific topic or problem
Systematic reviews assemble, critically appraise, and synthesize results of primary studies
Systematic reviews are prepared using explicit methods that limit bias and random error
Systematic reviews can help clinicians keep abreast of the overwhelming amount of medical literature
Systematic reviews can help predicate clinical decisions on research evidence
Systematic reviews are often more efficient and accurate than single studies

NOTE. Adapted from Cook et al.[91]

increasingly productive search for solutions to medical problems.[90] The features of a systematic review are listed in Table 13.

RATIONALE FOR CONDUCTING SYSTEMATIC REVIEWS

Quantity of Information

Over 2 million articles are published annually in the biomedical literature.[92] Decision makers of various types are inundated with an unmanageable amount of information. Systematic reviews are needed to refine this cumbersome amount of information. Practitioners and clinicians can use systematic reviews in place of an overwhelming volume of medical literature to keep informed.[91] In addition, through critical exploration, evaluation, and synthesis, systematic reviews are able to separate insignificant and unsound medical information from salient critical studies that should be incorporated into the clinical decision-making process.[92]

Integration

Systematic reviews integrating critical biomedical information are used by various decision makers. Research investigators need systematic reviews to summarize existing data, refine hypotheses, estimate sample sizes,[91] recognize and avoid pitfalls of previous investigations, and describe important secondary or adverse effects and covariates that may warrant consideration in future studies.[92] Without systematic reviews, researchers may miss promising leads or embark on studies inquiring into questions which have been previously answered.[91] Information encompassed within systematic reviews is also used by health policymakers to formulate guidelines and legislation regarding the use of certain diagnostic tools and treatment strategies as well as optimizing outcomes using available resources.[91,92] As previously discussed, systematic reviews are used by clinicians. Single studies rarely provide definitive answers to clinical questions. Systematic reviews can help practitioners solve specific clinical problems by ascertaining whether findings can be applied to specific subgroups, as well as keeping practitioners literate in broader aspects of medicine.[91,91] Lastly, systematic reviews shorten the time between medical research discoveries and clinical implementation of effective diagnostic or treatment strategies.[92]

Efficiency

Conducting a systematic review is usually more efficient, less costly, and quicker than embarking on a new study. It can also prevent pursuing research initiatives that have already been conducted.[92] Lastly, pooled results from various studies can give a better estimate of outcomes.

Generalizability

By using different eligibility criteria for participants, definitions of disease, methods of measuring exposure, sample sizes, populations, study designs, and variations of a treatment, multiple studies addressing the same question provide an interpretative context not available in any individual study.[92] Pooled results from these studies are more generalizable to the population than any individual study.[92]

Consistency

Systematic reviews can determine consistency among studies of the same intervention or among different interventions. Assessments of whether effects are in the same direction or of the same magnitude can also be made. Lastly, systematic reviews can help ascertain consistency of treatment effects across different diseases with a common underlying pathophysiology and consistency of risk factors across study populations.[92]

In addition to establishing consistencies, systematic reviews can be used to assess inconsistencies and conflicts in data.[92] Effectiveness of treatments in particular settings or only among certain subjects can be explored and assessed. Furthermore, findings from certain studies that stand alone because of uniqueness of the study population, study quality, or outcome measure can be explored.[92]

Increased Power and Precision: One of the most commonly cited reasons for conducting systematic reviews is the increase in power. Meta-analyses and pooled results yield increased statistical significance by increasing the sample size. The advantage of increasing power is particularly relevant to conditions of relatively low event rates or when small effects are being assessed.[92] Quantitative systematic reviews also allow for increased precision in estimates of risk or effect size. Meta-analyses show that increasing sample size from temporally consecutive studies results in a narrowing of confidence intervals.[92,93]

Accuracy: In contrast to traditional views, systematic reviews apply explicit scientific principles aimed at reducing random and systematic errors of bias and therefore lead to better and more accurate recommendations.[91] Furthermore, the use of explicit methods allows for an assessment of what was done and yields a better ability to replicate results in the future and understanding of why results and conclusions of reviews differ.

ELEMENTS OF A SYSTEMATIC REVIEW

A review is considered "systematic" if it is based on a clearly formulated question, identifies relevant studies, appraises the studies' quality, and summarizes evidence using an explicit and predetermined methodology (Table 13).

Step 1: Framing the Research Question

A good systematic review has a well-formed, clear question that meets the FINER (feasible, interesting, novel, ethical, and relevant) criteria.[94] Feasibility of the question is largely dependent on the existence of a set of studies that can be used to evaluate the question. The research question should describe the disease or condition of interest, the population, the intervention and comparison treatments, and the outcome(s) of interest.[94,95]

Step 2: Identifying Relevant Publications

Systematic reviews are based on a comprehensive and unbiased search of completed studies.[96] To capture as many relevant citations as possible, a wide range of medical, environmental, and scientific databases should be searched.[95] The Center for Review and Dissemination has compiled a comprehensive resource list for researchers undertaking systematic reviews.[97] The process for identifying studies to be included in the review and the sources for finding these studies should be established before conducting the review, such that they can be replicated by other investigators. Depending on the subject matter, MEDLINE, AIDSLINE, CINAHL, EMBASE, and CANCERLIT, among other databases, can be used. In addition, a manual review of the bibliographies of relevant published studies, previous reviews, evaluation of the Cochrane Collaboration database, and consultation with experts can also be undertaken.[94]

Criteria for Including and Excluding Studies: Before one conducts a systematic review, a rationale should be provided for including or excluding studies. Criteria for including or excluding studies typically specify the period in which the studies were published, the targeted population, the disease or condition of interest, the intervention of interest, acceptable control groups, an accepted length of loss to follow-up, required outcomes, and whether blinding should be in place. Though these are typical, other criteria can also be specified.[94] The criteria for inclusion and exclusion should be established before conducting the review.[95]

Once the criteria are established, each potentially eligible study should be reviewed for eligibility independently by 2 examiners. Any discrepancies should be settled by a third examiner or by consensus between the 2 examiners.[94] When determining eligibility, the examiners should be blinded to the dates of publication, authors of the study, and results to ensure an unbiased selection.[94]

Collecting Data From Eligible Studies: Predesigned forms should be created, which include variables such as eligibility criteria, design features, population included in the study, number of individuals in each group, intervention, and primary and secondary outcomes, as well as outcomes in subgroups.[94] The data should be abstracted individually by 2 independent assessors. As with the inclusion and exclusion of studies, if the 2 assessors disagree, a third assessor should settle the discrepancy or a consensus process may be used.[94]

Often, it is difficult to ascertain whether studies are eligible because published reports may or may not adequately describe important information such as design features, risk estimates, and standard deviations.[94] It is usually not appropriate to calculate risk estimates and confidence intervals based on crude data from observational studies because sufficient information may not be available for potential confounders. To attain adequate information, efforts should be made to contact the investigators and retrieve necessary information.[94]

Step 3: Assessing Study Quality

The greatest drawback to a systematic review is that the results can be no more reliable than the quality of the studies on which they are based.[94] If individual studies are of poor quality, this poses a significant risk to the overall quality of the systematic review. A simple procedure to ensure this is to create relatively strict criteria for good study design when establishing the inclusion and exclusion criteria. This is of particular importance when using observational studies. It is often difficult to conduct RCTs in evaluating public health interventions at the community level.[95] Therefore systematic reviews assessing the safety of such interventions need to include evidence from a broader range of study designs.[95] When using data from observational studies, results should be adjusted for potential confounding variables to ensure that results of meta-analyses are not confounded.[94]

Quality is a multidimensional concept that can relate to design, conduct, and analysis of the trial. Quality of a primary investigation can be affected by the presence of bias, which consequently affects internal validity. Assessing the quality of the studies included is currently debated.[98] Quality scores can combine information on several features in a single numerical value. Numerous quality checklists exist. However, caution must be exercised in their application because scores, and thus quality estimates, may differ across varying checklists. On the other hand, a component approach examines key dimensions individually.[98]

Incorporating study quality into meta-analysis can entail excluding trials that fail to meet some standard of quality. Although this may be justified, it can also lead to excluding studies that may contribute valid information.

Step 4: Meta-analysis—Summarizing the Evidence

Once all studies to be included have been identified and the data abstracted, a summary estimate and confidence interval may be calculated.[94] Methods for calculating the summary estimate and confidence interval, as well as principles of meta-analyses, are discussed in the next section. It is important to note that different approaches to calculating these estimates will yield different results.

Step 5: Presenting the Findings

Three types of information are typically included in systematic reviews. First, characteristics of each study are presented in tables. These often include study sample size, number of outcomes, length of follow-up, methods used in the study, and characteristics of the population studied. Second, results of individual studies are displayed. These can include risk estimates, confidence intervals, or *P* values.[94] Finally, the meta-analysis summary estimate, confidence interval, and subgroup and sensitivity analyses are presented. All information should be presented clearly in tables and figures.

META-ANALYSIS

Principles

After a systematic review, data from individual studies may be pooled quantitatively by use of established statistical methods. A useful definition of meta-analysis is given by Huque as "a statistical analysis that combines or integrates the results of several independent clinical trials considered by the analyst to be 'combinable.'"[90] The rationale for conducting meta-analysis is that combining individual studies provides an increased sample size, which improves the statistical power of the analysis and the precision of the estimates of treatment effects.[89]

Meta-analysis is a 2-stage process. First, it involves calculation of a measure of treatment effect with its 95% confidence interval for each individual study. This is accomplished by use of summary statistics such as odds ratios, relative risks, and risk differences. Second, an overall treatment effect is calculated as a weighted average of the individual summary statistics.[89] It should be noted that data from individual studies are not simply averaged. Instead, results are weighted. Higher weight is given to studies that provide more information.[89]

Heterogeneity

Combining the results of individual studies may not be appropriate if the results differ greatly. There are several ways to ascertain whether the results are heterogeneous and therefore inappropriate to combine.[99] First, individual studies can be reviewed to determine whether there are substantial differences in the study design, study population, interventions, or outcomes.[94] Second, the investigator can examine the results of individual studies; if some trials report a benefit whereas others report a significant harm, then the results are most likely heterogeneous. Statistical approaches exist to facilitate the establishment of heterogeneity in results.[94]

Tests of homogeneity assume that the results of individual studies are the same (the null hypothesis).

The test is used to determine whether the data refute this null hypothesis (the alternate hypothesis). A χ^2 test is commonly used. If the *P* value is greater than or equal to .10, then the data support the null hypothesis and the studies are homogeneous. If the *P* value is less than .10, then the null hypothesis is rejected and the study findings are considered to be heterogeneous. All meta-analyses should report the *P* value.[94,100]

If this test shows homogeneous results, then the differences between the studies can be attributed to sampling variation. In this case a fixed-effects model is used to combine the results. If the test indicates that heterogeneity exists between study results, then a random-effects model should be used to combine results.[99,100]

A major limitation to this approach is that statistical tests often lack power to reject the null hypothesis and studies appear to be homogeneous when they are not. There is no statistical solution to this problem.[94] Therefore a discussion of heterogeneity and its potential effects should always accompany summary estimates.[94,100]

Methods

Treatment Effects: The primary goal of meta-analysis is to calculate a summary effect size. If the outcome is binary (e.g., disease *v* no disease), then odds ratios or relative risks should be used. If the outcome is continuous (e.g., blood sugar measurement), then mean differences should be used.[89,100]

Odds ratio is defined as "the ratio of the odds of the treatment group to the odds of a control group."[89] Odds are calculated by dividing the number of patients in the group who achieve a certain endpoint by the number of patients who do not. Risk, in contrast to odds, is calculated as the number of patients in the group who achieve the stated endpoint divided by the total number of patients in the group.[89] Relative risk is the ratio of the 2 risks. An odds ratio or relative risk greater than 1 indicates increased likelihood of the stated outcome being achieved in the treatment group. Correspondingly, a relative risk or odds ratio of less than 1 indicates decreased likelihood of outcome in the treatment group. A ratio of 1 indicates no difference between the 2 groups. All estimates of relative risk and odds ratio should be accompanied by confidence intervals.[89]

Fixed- Versus Random-Effects Models: There are various statistical approaches to calculate a summary effect. These approaches are thoroughly discussed by Cooper and Hedges.[101] The choice of statistical method is dependent on the outcome measure and presence of heterogeneity. The fixed-effects model calculates the variance of a summary estimate based on the inverse of the sum of weights of each individual study.[94] The random-effect model adds variance to the summary effect in proportion to the variability of the results of the individual studies.[94] The confidence interval around the summary measure is usually greater in the random-effects model, and therefore the summary effects are less likely to be significant. Many journals now require authors to use the random-effects model because it is considered the most conservative.[94] It is also quite reasonable to use both a random- and fixed-effects model and present both estimates.

Confidence Intervals: Confidence intervals should accompany each summary measure. Intervals are commonly reported with 95% confidence but can be reported with 90% or 99% confidence.[89] A 95% confidence interval is the range within which the true treatment effect will lie with 95% certainty. The width of a confidence interval dictates precision; the wider the interval, the less the precision.

Many formulas exist to calculate the variance of summary risk estimates. The variance of the summary estimate is used to calculate the 95% confidence interval around the summary estimate ($\pm 1.96 \times \sqrt{\text{variance}}$).[94]

Assessment of Publication Bias

Publication bias occurs when published studies are not representative of all studies that have been conducted.[94] If reasons that studies remain unpublished are associated with their outcome, then meta-analyses combining the published results will be seriously biased. Hypothetically, with a treatment that has no actual effect on a disease of interest, studies that show a benefit may be published whereas studies that suggest harm may not be published. In this case a meta-analysis combining only published results would depict a beneficial impact.[102]

There are 2 main ways to circumvent the effects of publication bias. First, unpublished studies should be identified and included in the summary estimate. Unpublished studies can be identified by contacting investigators in the field and reviewing abstracts, meeting presentations, and doctoral theses. However, including unpublished studies can be problematic. It is often difficult to identify unpublished studies, and when identified, it is often difficult to extract relevant data, such as inclusion and exclusion criteria, or determine the quality of methods.[94] Efforts should be made, in these circumstances, to contact the investigators.

The extent of potential publication bias can be estimated. This estimate can then be reflected in the systematic review's conclusions. Publication bias exists when unpublished studies yield different results from pub-

lished studies. Unpublished studies are likely to be smaller than published studies and likely to have found no association between risk factor or intervention and the outcome of interest. If there is publication bias, there should exist an association between a study's sample size (or the variance of the outcome estimate; smaller studies tend to have larger variance) and findings. This association can be measured by use of the Kendall τ.[94] A strong correlation between sample size and findings would suggest a publication bias.[94]

Alternatively, a funnel plot can also be indicative of publication bias. In the absence of publication bias, a plot of the standard error versus log of outcome measure (i.e., odds ratio and relative risk) should have a funnel or bell shape (Fig 9A).[103] When publication bias is present, the plot is asymmetrical and truncated in a corner (Fig 9B).[103]

When substantial publication bias is present, summary estimates should not be calculated. If little publication bias is present, summary estimates should be interpreted with caution. All meta-analyses should contain a discussion of potential publication bias and its effect on the summary estimates presented.[94,102]

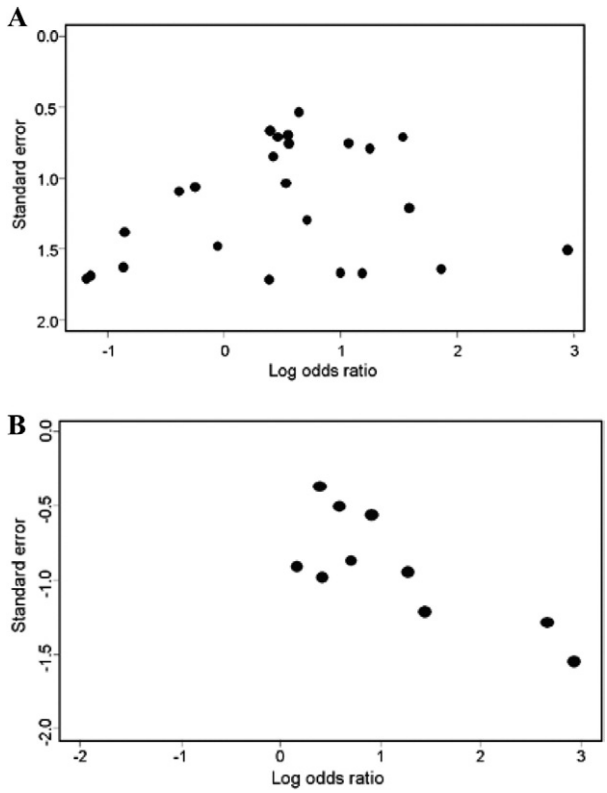

FIGURE 9. (A) Funnel plot that does not suggest publication bias.[22] (B) Funnel plot suggestive of publication bias.[22]

Subgroup and Sensitivity Analyses

Subgroup Analysis: The main aim of a meta-analysis is to produce an estimate of the average effect seen in trials of a particular treatment. The direction and magnitude of this overall effect are intended to guide clinical decision making. However, practitioners are presented with a problem when asked to use an average effect on specific groups of patients because the effect of a given treatment is likely to be different across different groups of patients.[104] It may, therefore, be possible to use data from all of the studies or some subset of the studies included in the systematic review.[94] Although meta-analyses offer a reliable basis for subgroup analyses, they are not exempt from bias and the results of such analyses should always be interpreted with caution.[104]

Sensitivity Analysis: The robustness of findings of a meta-analysis should be examined through sensitivity analyses.[100] An analysis can entail an assessment of the influence of methodologic quality and the presence of publication bias.[102] Quality summary scores or categorical data on individual components can be used to explore the methodologic quality. Simple stratified analyses and meta-regression models are useful for exploring associations between outcome effects and study characteristics.[98]

IMPLEMENTATION AND COCHRANE COLLABORATION

Despite the considerable amount of resources spent on clinical research, relatively little attention has been given to ensuring that the findings of research are implemented in routine clinical practice.[105] There are many strategies for intervention that can be used to promote behavioral changes among health practitioners (Table 14).

The Cochrane Collaboration, an international organization, has facilitated informed decision making in health care by preparing, maintaining, and promoting the accessibility of systematic reviews on the effects of health care interventions.[106] These reviews are available in the Cochrane handbook, updated and modified in response to new evidence.[106] Because Cochrane reviews have greater methodologic rigor and are more frequently updated than systematic reviews published in paper-based journals, they present an excellent resource to be used in clinical practice.

CONCLUSIONS

Systematic reviews are scientific investigations conducted with a specific methodology using independent

TABLE 14. *Interventions to Promote Behavioral Change Among Health Professionals*[105]

Consistently effective interventions
Educational outreach visits
Computerized or manual (e.g., mail) reminders
Multifaceted interventions (a combination of audit and feedback, reminders, local consensus processes, and marketing)
Interactive educational meetings (participation of health care practitioners in seminars or workshops that include discussions)
Interventions of variable effectiveness
Audit and feedback (e.g., summary of clinical performance)
Local consensus processes (focus groups, discussion with experts and local practitioners)
Patient-mediated interventions (aimed at changing performance of health care providers)
Interventions that have little or no effect
Educational materials (distribution of information pamphlets or best practice guidelines)
Didactic educational meetings

NOTE. Adapted from Bero et al.[105]

studies. They appraise study characteristics, synthesize the results of multiple primary investigations using established strategies, and interpret results in a manner aimed at limiting bias. Strategies include a comprehensive search of relevant articles using explicitly defined and reproducible criteria. A detailed explanation of the steps involved in conducting a systematic review was discussed in this chapter. Systematic reviews often use meta-analyses to combine results of the eligible studies to increase sample size, which improves the statistical power of the analysis and the precision of the estimates of treatment effects. Meta-analysis is a 2-stage process, which involves calculating a treatment effect with its 95% confidence interval for each individual study and, if appropriate, calculating an overall treatment effect as a weighted average of the individual summary statistics. The specifics of conducting a meta-analysis were also discussed.

Zahra N. Sohani, M.Sc.
Jón Karlsson, M.D., Ph.D.
Mohit Bhandari, M.D., Ph.D., F.R.C.S.C.

SECTION 9

Special Designs: Survey and Reliability Studies

Survey and reliability studies are a valuable element of outcomes research. With survey studies, it is very important to determine which questions to ask. Each question should be viewed as a possible measure of general or joint-specific health, and each data point collected should be potentially valuable in addressing the research question about how the patient is progressing.[107] This is an important first step to addressing what questions should be on future survey studies so that these surveys can contribute valuable information to patient assessment. The psychometric properties of a survey should also be established.[107-110] This will determine whether the questions are valid, reliable, and responsive. If a survey is not tested for these parameters, it will be unclear whether the survey is measuring what it is supposed to measure, whether it is accurate, and whether it can measure change. Without these, the results of the survey may come under question.

Reliability studies help determine what are accurate and consistent measurements and whether these measures can be consistently interpreted.[111,112] These studies are commonly done on radiographic measurements. The reliability of a radiographic measurement allows for clinicians to compare their measurement with those of other centers given the measurement reliability.

SURVEY STUDIES

When approaching a topic of study, it is essential to keep in mind your research purpose and the specific questions that you are trying to answer with the use of your research instrument. Each question on the instrument should be a measure that addresses your research questions. If you are asking a question that is not useful in terms of answering the research questions, then it should be removed. The only exceptions would be information collected to control for population factors such as gender, age, smoking status, and so on. Too many questions on an instrument may interfere

with obtaining complete responses and thereby interfere with the research purpose.

Before embarking on designing a new questionnaire, a thorough review of the literature is important to determine what scales are currently being used and whether your question has been previously addressed. This is also helpful in determining what the best method of addressing the question might be, namely what are the unanswered questions in the field of study, what is known, what are the problems with current research, and what would be clinically useful.

The usual method of administration in orthopaedics is by questionnaire. Questionnaires are less expensive, can be standardized, and can measure subjective data at different points in time to determine the outcomes of an intervention.[113] The design of the questions on your study instrument is important for collection of useful data. If you are using a standardized, scannable questionnaire that will be entered into a database, it is necessary to consider the fields on the form and how the data will be analyzed. For example, if you are asking a question about the types of symptoms that a patient has, he or she is given a list of 10 possible symptoms, all of these symptoms are entered into 1 field, and then these data will have to be re-coded for use by the analysis software. In addition, the lack of a response for any given symptom may indicate that patients do not have the symptom or they did not answer the question, even though they do have this symptom. However, if you ask the question as, "Do you have the following symptoms?" and each symptom listed is a field with a possible yes or no response, then these data will not have to be re-coded for analysis and it is clear whether patients have the symptom or not if they mark yes or no or if they missed the question and did not answer yes or no.

The wording of the questions is also important to obtain complete and meaningful answers. If the respondent does not understand the question, he or she might not answer the question. Worse, if the respondent misunderstands the question and answers it incorrectly, then the data might be skewed and your results will not make sense and/or be useful. Questions should also be asked in a neutral manner. It is important to avoid words of judgment. For example, the question "On average, how many days per week do you abuse alcohol?" prejudges the respondent's answer, whereas the question "On average, how many days per week do you drink more than 4 alcoholic drinks?" allows the respondent to answer with less judgment. The wording should allow respondents to answer honestly but not encourage them to exaggerate. The wording should also be at a reasonable educational level that takes into consideration the spectrum of your study population. Most often, a sixth-grade reading level will allow for sufficient information to be collected but still be accessible to most of the respondents.

TABLE 15. *Keys to Survey Research*

1. Questionnaires should be designed so that they are easy to read, understand, and complete. Poorly designed questionnaires will lead to incomplete data.
2. Questions and outcome scores should be reliable, valid, and responsive.
3. Define a clear data collection protocol. A good operating procedure will improve the quality of the data collection and the data.

The format of your questionnaire can help the respondent to complete the forms (Table 15). For example, if you have a question that asks the respondent to rate limitation on a 1 to 10 scale followed by a question that asks the respondent to rate activity on a 0 to 10 scale, it might be confusing whether 10 represents the most limitation or the most activity. It is better to clearly separate these questions to visually cue the respondent to recognize that 10 represents the most activity in one case and the most limitation in another. In addition, the order of the questions is important to help the respondent feel comfortable with the study. Questions about personal or potentially embarrassing information should come later in the questionnaire and preferably not on the first page. For example, questions about income, sex, and/or recreational or prescription drug use should not be at the beginning of the questionnaire. Finally, the font should be large enough for visually impaired respondents to see, and answers should be visually consistent so as not to frustrate the respondent or make him or her dizzy.

The use of outcome instruments whose psychometric properties have been vigorously established is essential. Psychometric evaluation is the epidemiologic field that studies the development and validation of outcome instruments and questionnaires.[107-110,114] The important psychometric properties of an outcome instrument include reliability, validity, and responsiveness.[107-110] Reliability refers to the reproducibility of an outcome measure, either between subjects (test-retest reliability) or between observers (interobserver reliability). Validity questions whether an outcome instrument actually measures what it intends to measure. Components of validity include content validity

("face" validity and floor/ceiling effects), criterion validity (how an instrument compares to an accepted "gold standard" instrument), and construct validity (whether the instrument follows expected noncontroversial hypotheses). Responsiveness assesses change in the instrument's value over time or treatment.

Test-retest reliability is determined by comparing an original questionnaire and a second postoperative questionnaire given in a short time span when no change in clinical status has occurred. The intraclass correlation coefficient (ICC) is determined for each component score. An ICC greater than 0.70 is considered acceptable.[110] The standard error of the measurement is also calculated as described previously.[115] This value will be used to determine the 95% confidence interval for individual scores, which provides an estimate of where the actual score may lie. To further define this interval, the minimum detectable change is calculated to determine the smallest change that can be considered a true difference after measurement error and noise have been taken into account.[115] Noise would be changes in the score due to factors other than changes due to the intervention.

Content validity is determined by the floor and ceiling effect of the score. Preoperative scores are used to establish content validity. Floor effects (scale = lowest possible) and ceiling effects (scale = highest possible) will be determined for each component. Floor and ceiling effects of less than 30% were considered acceptable.[110]

Criterion validity is determined by the correlation of the score with a gold standard. The definition of a gold standard is a score that has been validated for the population you are studying. It is common in orthopaedics to use the Short Form (SF)-12 or SF-36 as a gold standard because it has been extensively studied.[116] The Pearson correlation coefficient should be used for the continuous variables that are normally distributed and the Spearman ρ should be used for nonparametric data.

Construct validity is determined by developing 5 to 10 hypotheses or constructs that are noncontroversial and considered true by many surgeons—for example, "patients with severe pain have lower activity level." These constructs are developed by consensus and tested in the study population. Construct validity tests the score to make sure that score can measure what it claims to measure. If it is a functional score, then it should pass tests that are considered true differences in function.

Responsiveness to change is assessed by comparing the initial scores with scores after an intervention. The time between the initial and follow-up scores should be long enough for the intervention to have made a difference. For example, you would not measure function after ACL reconstruction 2 days after the surgery. Effect size is calculated as (mean postoperative scale – mean preoperative scale)/standard deviation of preoperative scale. Standardized response mean is calculated as (mean postoperative scale – mean preoperative scale)/standard deviation of change in scale. Small effects are considered greater than 0.20, moderate effects are considered greater than 0.50, and large effects are considered greater than 0.80.[110]

The data collection protocol should include the times at which each measure will be collected, with clear specifications for ranges of acceptable collection times. For example, if an outcomes measure will be collected before surgery and at 1 year and 2 years after surgery, it should be clearly stated that the measure can be collected within 1 month before surgery and within 1 month before or after the 1-year and 2-year marks. The collection protocol should also detail who will collect the data. If the data are being collected by 1 or more observers, then the protocol should detail what their qualifications are and how they will collect the data. If any tools are used to collect the data, then they should be described and the methodology for collection should be described. If calculations are required, the method of calculation should be described, especially if a specific type of software is being used.

RELIABILITY STUDIES

Reliability of a measurement defines whether the measurement is dependable and reproducible. Reliability studies answer the following questions: (1) Is this measurement consistent? (2) Is this measurement free of error? Intraobserver reliability and interobserver reliability compare the scoring of tests by the same observer and by different observers, respectively.[111] Intraobserver reliability tests the ability of 1 observer to duplicate his or her test responses on the same test at different points in time, when no intervention for the disease has taken place and/or there has been no progression. Interobserver reliability tests the ability of more than 1 observer to give similar responses on the same test. Reliability is a measure of reproducibility, not of accuracy, and the different ways to measure reliability each provide insight into reproducibility.

When initiating a study that included objective measurements, such as alpha angle in the hip, it is important to include a study of the reliability of the mea-

surement (Table 16). Although it may be quoted in the literature, it is important for readers to know the reliability of the measurement in individual practice settings. When designing the reliability arm of the study, it is important to consider who will be the observers in the study. It is assumed that an observer with more training—for example, a senior physician with 10 years of surgical experience—would be more reliable in measuring than a resident. If the measurement you are testing is commonly measured by residents and senior physicians, it will be important to include both in your study.

When identifying the group of patients to be used in the reliability study, you must ensure that the group covers all levels of the scores. For example, if you are testing the reliability of scoring Kellgren-Lawrence grades on knee radiographs, you do not want to only include those with severe OA. This will make it easier for the observers. Make sure there are a number of patients with every grade in the study group.

After the group has been identified, the object to be tested, for example, the anteroposterior radiograph of the knee, should be de-identified. It should be numbered and any information on the patient should be removed, including his or her name. If the patient has been examined by the observer, the observer should not know that the patient is in the reliability study. Using a random-number generator, you can determine the order in which the radiographs will be observed. If you have 20 radiographs, then use a random-number generator between 1 and 20, and this will provide you with the order of the first reading. For the second reading, again randomize the order for the samples.

Determining the number of observers is usually based on clinical practicality. To have 5 different observers go to a patient's room and measure his or her hip range of motion may be impractical for the patient and the clinic. To determine the sample size needed for the number of subjects, the acceptable level of reliability for the measure must be known.[111,117] In addition, sample size may be different depending on whether the measurement represents continuous or categorical data.[117]

TABLE 16. *Inter-Rater Versus Intrarater Reliability*

Inter-rater reliability of a measure will define whether the agreement between 2 observers is acceptable.
Intrarater reliability of a measure will define whether there is agreement between 2 observations by the same observer at different times.

TABLE 17. *Survey Studies*

1. Survey studies should be based on clinically relevant questions.
2. Psychometric properties of outcome scores are important and should be determined before using the score.
3. Important information can be obtained through survey research. The data represent a valuable research tool, and this also allows physicians to track patients and improve patient care.

For continuous variables, such as degrees, inches, and so on, the ICC is used to measure reliability. The ICC is a ratio of the variability among subjects to the overall variability observed in the data. A score of 0 to 0.4 indicates poor reliability, a score of greater than 0.4 to 0.75 indicates fair or moderate reliability, and a score of greater than 0.75 indicates excellent reliability.

For categorical data, the κ coefficient is used to report reliability. The κ coefficient measures the observed agreement compared with the possible agreement beyond chance. For more complicated models, a statistician should be consulted.

CONCLUSIONS

The goal of reliability and survey studies is to measure patient health accurately (Table 17). Specifically, information about how patient health will be affected and how long it will be sustained are essential factors for improving patient care in the future. The key to obtaining this valuable knowledge is good measurement grounded in an understanding of what and how health is being measured.[110] Reliability, validity, and responsiveness studies of disease-specific outcomes instruments provide researchers the tools they need to make accurate measurements of patient health. Survey studies that use these outcomes measures can provide surgeons with the clinically relevant patient information they need to improve function and activity levels for patients with varying types of orthopaedic disease. Survey and reliability studies are therefore valuable tools in the process of continually improving patient care.

Karen K. Briggs, M.P.H.
Kira Chaney-Barclay, M.P.H.
Robert F. LaPrade, M.D., Ph.D.

SECTION 10

Outcome Measures: A Primer

In orthopaedic surgery, new technologic developments and advances are common. However, the introduction of and acceptance of these new developments must be guided by appropriate levels of evidence. It follows that these new technologies should be compared with current technologies (gold standard) in well-designed trials. To ensure patient safety, decisions such as using new devices must be based on the best available evidence.

A well-designed, blinded, prospective RCT is one of the best ways to provide credible evidence. By concealing allocation of treatment, randomly allocating treatment groups, and blinding the outcome observers and patients, bias is limited.[118] Thus a novel intervention can be tested against the current standard accurately for an outcome in question (pain, range of motion, outcome scores, and so on). A study design following these principles can give answers that can be readily applied in clinical practice.

CHOOSING THE RIGHT OUTCOME

During the early stages of study design, choosing an appropriate outcome measure is critical. Instruments of measure are considered useful in assessing orthopaedic outcomes if they are valid, reliable, and responsive to change.

Validity

The validity of an instrument refers to its ability to measure what it is supposed to measure. The term "validity" consists of several types, including face validity, content validity, construct validity, convergent validity, and predictive validity.

Face validity refers to how well the items or questions represent the construct that is being measured. For example, a measure of knee pain would have sufficient face validity if the items on the measuring instrument ask the patient about specifics relating to knee pain. This is a very rudimentary type of validity.

Content validity refers to whether the items that make up the scale include all the relevant aspects of the construct that is supposed to be measured. For example, in patients who undergo shoulder arthroplasty and then are assessed with the Disabilities of the Arm, Shoulder and Hand (DASH) questionnaire, the content validity of the questionnaire would be whether the questionnaire includes questions relevant to all aspects pertaining to shoulder pain and function.

Construct validity refers to the theoretical framework of a general concept or idea. Such a concept would be the overall health of an individual, which will include physical, social, and emotional health. The general health questionnaire, developed by Sir David Goldberg, is considered to have excellent construct validity. In the study by Wright and Young,[119] the construct validity of the Patient-Specific Index was evaluated by comparing the scores obtained with those of the Harris Hip Score, the Western Ontario and McMaster University Osteoarthritis Index (WOMAC), the McMaster-Toronto Arthritis Patient Preference Disability Questionnaire, and the SF-36.

Convergent validity pertains to whether scores of a particular measure correlate with scores of other measures that measure the same construct. "For example, one would expect two instruments that claim to measure quality of life in patients with osteoarthritis of the knee to behave in a similar manner in response to arthroplasty."[120] In contrast, discriminant validity pertains to a situation in which the scores on a particular measure are not correlated with scores on other measures that assess an unrelated construct.

Predictive validity refers to whether the score of a measure can predict a patient's score on a measure of some related construct.

Validity in Action—A Case Example From the Literature: To evaluate the validity of an outcome measure, the results should be compared with a "gold standard" to ensure that the measurement tool is measuring what it is supposed to measure. In the absence of a gold standard, investigators rely on construct validation correlating the baseline scores with change scores in their own scale. These values are then compared with other scales measuring the same/similar outcomes, and if the prediction of how the tool relates to other measures is confirmed in the population of interest, the evidence for validity is strengthened.

To illustrate this concept further, we will discuss how the Western Ontario Shoulder Instability Index (WOSI), a 21-item disease-specific quality-of-life

measurement tool for shoulder instability developed by Kirkley et al.,[121] was validated. Because there was no "gold standard" for quality of life, Kirkley et al. used construct validation to demonstrate how the WOSI "behaved" in relation to 5 other measures of shoulder function. They administered the WOSI on 2 occasions to a randomly selected group of 47 patients undergoing treatment for shoulder instability. Also administered to these same patients were the DASH measurement tool; the Constant score; the University of California, Los Angeles (UCLA) Shoulder Rating Scale; the American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form; the Rowe Rating Scale; the SF-12 global health instrument; shoulder range-of-motion evaluation; and a global change rating scale.[121] The correlations of the baseline and change scores were determined by the Pearson product-moment correlation. The WOSI correlated highly with the DASH as well as with the UCLA score, perhaps reflecting the importance patients place on pain.

Reliability

Reliability of an instrument refers to the consistency with which a given outcome occurs, given repeated administrations of the test. Many aspects of reliability can be assessed: intrarater, inter-rater, test/retest, and internal consistency reliability.

Intrarater reliability is defined as the agreement between scores of one rater's 2 or more assessments at different time periods. Inter-rater reliability is defined as the agreement between scores of 2 or more raters' assessments. Test-retest reliability is the agreement between observations on the same patients on 2 or more occasions separated by a time interval under stable health conditions.

Reliability in Action—A Case Example From the Literature: Wright and Young[119] developed the Patient-Specific Index, which consisted of ratings of the importance and severity of several concerns of patients scheduled for hip arthroplasty. In a well-designed RCT, the Patient-Specific Index was administered before patients underwent total hip arthroplasty and 6 months later to determine the reliability, validity, and responsiveness of this scale. The test-retest reliability of the Patient-Specific Index was determined by interviewing 30 patients twice, 2 weeks apart, before the operation. The choice of 2 weeks was based on the thinking that the patients would not remember their previous responses and that their clinical status would remain constant. The sample size calculation was based on the random-effects ICC.

The ICC is one of the statistical measures that can be used to quantify test-retest reliability over time, i.e., the extent to which the same test results are obtained for repeated assessments when no real change occurs in the intervening period. The ICC can range from 0.00 (no agreement) to 1.00 (perfect agreement).[122] An ICC equal to or greater than 0.70 can be regarded as adequate for group comparisons, and an ICC equal to or greater than 0.90 is required for a reliable assessment of an individual.[123]

With athletic patients as their subjects, Marx et al.[124] evaluated the reliability, validity, and responsiveness to change of 4 rating scales for disorders of the knee: the Lysholm scale, the subjective components of the Cincinnati Knee Rating System, the American Academy of Orthopaedic Surgeons sports knee-rating scale, and the Activities of Daily Living scale of the Knee Outcome Survey. Forty-one patients who had a knee disorder that had stabilized and who were not receiving treatment were administered all 4 questionnaires at baseline and again at a mean of 5.2 days (range, 2 to 14 days) later to test reliability.[124] The ICC was the mathematical measure used to compare the scores.

The reliability of all 4 scales was excellent. The ICC was 0.88 for the Cincinnati Knee Rating System, 0.95 for the Lysholm scale, 0.93 for the Activities of Daily Living scale, and 0.92 for the American Academy of Orthopaedic Surgeons sports knee-rating scale. Therefore all 4 scales are adequate for patients enrolled in a clinical trial and considered reliable.[124]

Instrument reliability, or internal consistency, can be evaluated with the ICC or Cronbach α.

Internal consistency can be quantified by the average correlation among questions or items in an outcome measure or scale and is expressed as the Cronbach α. To quantify the internal consistency of items within a scale, the Cronbach α is used; it ranges from a value of 0.00, representing no correlation, to 1.00, representing perfect correlation. The questionnaire would be considered to be internally consistent if the Cronbach α was between 0.7 and 0.9; thus a Cronbach α of 0.8 is considered good, and a value of 0.9 is excellent. However, a value greater than 0.9 is too high and represents an outcome scale in which many items are likely measuring the same aspect twice.

Several factors influence reliability of a measure between test dates, including "differences between the conditions of administration, the effects caused by repeated testing, such as learning and regression to the mean, factors affecting participants in their daily lives, and the length of time between administrations."[120]

Using Common Sense for Your Population When Testing Reliability: When testing the reliability of a scale, the population in which it is tested is important, e.g., when assessing a scale on ACL stability, if most patients have stable ACLs, the inter-rater reliability will be very high and there will be very little disagreement, giving a false impression of a scale with high inter-rater reliability. The patient population should consist of patients whose ACLs range between stable, mildly stable, and unstable. Then, if the inter-rater reliability is high, it is much more likely to be a true value.

Responsiveness to Change

Responsiveness to change is the ability of an instrument to detect clinically important changes between the patient's pre-intervention and post-intervention state, assuming all other factors are held constant. For example, Dias et al.[125] assessed the responsiveness of their Patient Evaluation Measure in detecting clinically important changes in pain, tenderness, swelling, wrist movement, and grip strength in patients with a scaphoid fracture.

For evaluative instruments designed to measure longitudinal change over time, the instrument must detect clinically important changes over time, even if small. In the study by Wright and Young,[119] responsiveness was assessed by measuring the change in the patient's mean severity-importance of his or her complaints from the preoperative period to the postoperative period. On average, the severity-importance values improved for practically all complaints. To test whether the values were responsive to change, the responsiveness statistic was calculated as the ratio of the clinical change after a known therapeutic intervention divided by the variability in test scores for stable subjects.

Many statistics are available to determine responsiveness. Another method used in orthopaedic surgery is the "standardized response mean," which is the mean change in score divided by the standard deviation of the change scores; it has been used by Kirkley et al.[121] and Marx et al.[124]

MEASURING QUALITY OF LIFE

According to the World Health Organization, health is "a state of complete physical, mental, and social well-being."[126] Outcomes research seeks to provide patients with knowledge regarding their expected functional recovery, as well as psychological and social well-being, and the delivery of their care from information obtained by studying the end results of surgical practices and interventions that directly affect both the patient and the global health care environment.

The patient's own assessment of outcomes in orthopaedic surgery is especially important. Thus outcomes research should take into consideration patient perspectives in judging the results of a treatment.

TRADITIONAL OUTCOME MEASURES

Traditionally, clinical outcome measures in orthopaedic surgery consisted of measuring impairments, such as range-of-motion and strength impairments, as well as pain.[127,128] Surgeons were not as interested in the functional limitations and disability, but because these are important to the patient, surgeons should quantify their dysfunction. The patient's perception of changes in health status is the most important indicator of the success of a treatment. Accordingly, patients' reports of function have become important outcome measures.[124,127,129] These measures allow clinicians to measure changes in functional limitations and disabilities after surgical interventions. An example is the Foot and Ankle Disability Index (FADI), which was designed to assess functional limitations related to foot and ankle conditions.[127,130]

WHAT IS THE APPROPRIATE OUTCOME MEASURE PERSPECTIVE?

The selection of outcome measures should concern the surgeon, the hospital, the payer (patient, insurance, government, and so on), and society, but most importantly, it should focus on the patient and the outcomes that are important from his or her viewpoint.

Outcome instrument development usually begins with qualitative research, using focus groups or qualitative interviews. Focus groups consist of groups of people who discuss their attitudes, interests, and priorities toward the research topic. Interaction and conversation are facilitated with structured questions for the group. This approach to instrument development can be used in knee surgery, for example, to better understand OA patients' physical limitations, physical priorities, and concerns with medical and surgical treatment options.[131]

The qualitative information gathered from the focus groups is used to form the conceptual model, from which the questionnaire is developed.[131] The questionnaire's validity, reliability, and responsiveness to change should be tested. Finally, the questionnaire should be feasible to apply.[131]

HEALTH-RELATED QUALITY-OF-LIFE MEASURES

The 2 ways of measuring health-related quality of life (HRQL) are measuring health from a broad perspective, called "generic measures," and measuring relative to a specific problem or function, called "disease-specific measures."

Generic measures pertain to the overall health of the patient, including physical, mental, and social well-being. With generic measures, such as the SF-36, NHP (Nottingham Health Profile), and SIP (Sickness Impact Profile), overall health states can be compared before and after an orthopaedic procedure. Advantages of generic measures include their breadth, scope, and comparative value because they can be used to compare health states across different diseases, severities, and interventions and, in some cases, across different cultures. The disadvantage is that generic measures may not be sensitive enough to detect small but important changes and have too wide of a focus to be used in subspecialty disciplines.

Disease-specific measures pertain to a specific pathology treated in a patient. These measure the specific physical, mental, and social aspects of health affected by the disease (e.g., WOMAC for knee and hip OA, DASH, NULI [Neck and Upper Limb Index], and MHQ [Michigan Hand Outcomes Questionnaire]). The greatest advantage of disease-specific measures is detecting small but important changes. The disadvantages are that they are not generalizable and that they cannot compare health states across different diseases.

For the purpose of providing a complete picture of the effect of a treatment on a patient, the patient should be assessed with a disease-specific measure (e.g., WOMAC) in combination with a generic measure (e.g., SF-36) to provide a complete picture of the effect of a treatment on a patient. If possible, the investigator should consider the use of a utility measure, which is an outcome measure pertaining to cost analysis.

In outcomes research, endpoint measures "and the instruments used to evaluate these endpoints are often disease or region specific. Investigators are challenged to use appropriate techniques to measure common endpoints, such as HRQL, economic burden, and patient satisfaction, in a reliable and valid manner across multiple health conditions."[131]

Using HRQL data can give a patient a way to compare his or her options based on the experience of previous patients who underwent the same procedure or a similar procedure. For example, a radiograph of the knee will not provide much insight into the patient's overall health state; however, generic health outcomes with patient satisfaction and HRQL data provide this information. Subsequently, a patient can decide whether the perioperative risks and acute pain from a hemiarthroplasty of the knee will be worthwhile, given the decreased long-term pain and increased knee range of motion.

CONCLUSIONS

Outcome measures should focus on what is important to the patient. When evaluating an outcome, a disease-specific measure should be used in conjunction with a generic measure, and if possible, HRQL data can provide a tangible way for the physician to present patients with information on what overall impact undergoing treatment may have on their quality of life. Finally, to choose the correct outcome measure, surgeons need to be able to evaluate the quality and usefulness of an outcome measure for a specific disease state.

Sophocles Voineskos, M.D.
Olufemi R. Ayeni, M.D., F.R.C.S.C.
Mohit Bhandari, M.D., Ph.D., F.R.C.S.C.

SECTION 11

Common Scales and Checklists in Sports Medicine Research

The improvement of surgical outcomes for patients in the future requires the evaluation and comparison of surgical outcomes of patients from the past. This is a principle behind much clinical research that has driven the development of instruments to make this kind of evaluation and comparison possible. Rat-

ing scales are particularly useful with regard to comparison: whereas the words of one surgeon's subjective description of an outcome may not be comparable to another surgeon's description, numbers can be compared easily. However, that comparison is meaningful only if the numbers are produced by rating scales that are reliable, valid, and responsive.[132] That is, the rating scale must be precise, it must prove accurate, and it must remain precise and accurate even as the patient's outcome changes over time.

Reliability could also be termed reproducibility, and there are 2 ways of evaluating it.[133] Test-retest reliability measures consistency over time. Patients whose clinical states are not expected to change are asked to take the test at 2 points in time, and the scores are compared. The time interval between tests is chosen so that patients will neither have experienced a change in their clinical state nor remember their previous responses.[134] Reliability can also be measured in terms of internal consistency, borrowing psychometric concepts to arrive at a statistic representing the inter-correlation of a patient's responses to questions on one administration of the rating scale questionnaire (usually Cronbach's α).[132]

A valid instrument is one that measures what it aims to measure. Criterion validity is the most straightforward: comparison of the rating scale results to a gold standard.[135] This is generally impossible for HRQL. Face validity is more subjective, consisting of the belief of an expert clinician that the instrument does indeed measure the concept in question. Content validity conceptually formalizes face validity and is present when the instrument measures the components of the overarching concept in question. Construct validity involves comparison of the instrument's results with those of other instruments, with which the instrument in question would be expected to correlate positively or negatively.[132] Responsiveness, or sensitivity to change in outcome, is necessary for the practical application of an outcome rating scale, because clinicians are especially interested in facilitating and measuring patients' improvement over time.[132]

This section reviews the reliability, validity, and responsiveness of outcome rating scales of the shoulder, knee, and ankle. Generally, studies should pair disease- or anatomy-specific scales like these with general outcomes measures to make comprehensive evaluation and cross-disease comparison of conditions possible.[136]

SHOULDER RATING SCALES

Many scoring systems have been developed to measure the clinical status and quality of life in patients with different pathologies of the shoulder. Initially, scales were developed when little information was available on the appropriate methodology for instrument development. Today, an appropriate instrument exists for each of the main conditions of the shoulder. Investigators planning clinical trials should select modern instruments that have been developed with appropriate patient input for item generation and reduction, as well as established validity and reliability.[137] In addition, the responsiveness of a scoring system is an important consideration because it can serve to minimize the sample size for a proposed study. We will present the most commonly used shoulder scales (Table 18), commenting on their strengths and weaknesses.

Clinician-Based Outcome Scales

In 1978 Rowe et al.[138] published the well-known rating system for the postoperative assessment of patients undergoing Bankart repair surgery: the rating sheet for Bankart repair (already known as the Rowe score). This system was very simple and based on 3

TABLE 18. *Shoulder Rating Scales*

	Instability	Rotator Cuff Disease	Osteoarthritis	Global Evaluation
Clinician-based outcome scales	Rowe (1978) UCLA (1981) ASES (1993)	UCLA (1981) Constant (1987) ASES (1993)	UCLA (1981) Constant (1987) ASES (1993)	UCLA (1981) ASES (1993)
Patient-related outcome scales	Oxford shoulder instability questionnaire (1999) WOSI (1998)	Rotator cuff quality of life (2000) WORC (2003)	Oxford shoulder score (1996) WOOS (2001)	Shoulder rating questionnaire (1997) DASH (1996-2002)*

NOTE. Scales are listed in increasing order of validity and reliability.

Abbreviations: ASES, American Shoulder and Elbow Surgeons; WORC, Western Ontario Rotator Cuff; WOOS, Western Ontario Osteoarthritis of the Shoulder.

*DASH is an outcome tool to be used for patients with any condition of any joint of the upper extremity.

separate areas: stability accounts for 50 points, motion for 20 points, and function for 30 points, giving a total possible score of 100 points.

In 1981 Amstutz et al.[139] introduced a rating scale intended to be used in studies of patients undergoing total shoulder arthroplasty for arthritis of the shoulder: the UCLA shoulder rating scale. Since then, however, it has been used for patients with other shoulder pathologies including rotator cuff disease[140] and shoulder instability.[141] This instrument assigns a score to patients based on 5 separate domains with different weights: pain, 28.6%; function, 28.6%; range of motion, 14.3%; strength, 14.3%; and satisfaction, 14.3%. There is 1 item for each of these areas, giving a total of 35 points.

The Constant score,[142] introduced in 1987, combines physical examination tests with subjective evaluations by the patients. The subjective assessment consists of 35 points, and the remaining 65 points are assigned for the physical examination assessment. The subjective assessment includes a single item for pain (15 points) and 4 items for activities of daily living (work, 4 points; sport, 4 points; sleep, 2 points; and positioning the hand in space, 10 points). The objective assessment includes range of motion (forward elevation, 10 points; lateral elevation, 10 points; internal rotation, 10 points; and external rotation, 10 points) and power (scoring based on the number of pounds of pull the patient can resist in abduction to a maximum of 25 points). The total possible score is therefore 100 points. The strength of this instrument is that the method for administering the tool is quite clearly described, which is an improvement on preexisting tools. This instrument is weighted heavily on range of motion (40%) and strength (25%). Although this may be useful for discriminating between patients with significant rotator cuff disease or OA, it is not useful for patients with instability.

There are many problems that can be identified with the previously described rating systems (Rowe, UCLA, and Constant scores). There are no published reports on the development of these instruments. It is likely that items used in the questionnaires were selected without direct patient input. It is unknown why the developers assigned different weights to the various items. Although this is not necessarily incorrect, it is unsupported. Some physical examination tests are not well-described in the first 2 scales (Rowe and UCLA scores). Moreover, these instruments combine items of subjective evaluation with items of physical examination for a total score: because these items are measuring different attributes, it is not ideal to combine them. Only the reliability of the Constant score has been evaluated. Conboy et al.[143] measured the reliability in 25 patients with varying shoulder syndromes, showing that the 95% confidence limit was 27.7 points between observers and 16 points within observers. Otherwise, no data on the formal testing of validity or the responsiveness of these instruments have been published.

All of these scores were developed before the advent of modern measurement methodology. The problems identified with these tests may lead to poor reliability, validity, and responsiveness, and therefore they may or may not be ideal choices for research, because they may not reflect what matters most to patients.[137]

In 1993 the American Shoulder and Elbow Surgeons developed a standardized form (the ASES score) for the assessment of shoulder function.[144] The instrument consists of 2 sections. The physician-assessment section includes physical examination and documentation of range of motion, strength, and instability, as well as demonstration of specific physical signs; no score is derived for this section of the instrument. The patient self-evaluation section has 11 items that can be used to generate a score, divided into 2 areas: pain (1 item) and function (10 items). The final score is tabulated by multiplying the pain score (maximum, 10) by 5 (thus a total possible of 50) and the cumulative activity score (maximum, 30) by 5/3 (thus a total possible of 50), for a total of 100. No rationale has been presented for the weighting of this instrument. Though not necessarily incorrect, it is unsupported. No data are available in the current literature on the testing of this instrument. The first developed shoulder scales reviewed so far can be used to investigate different shoulder pathologies (Table 18).[141]

Patient-Related Outcome Scales

In the last 15 years the need for a well-accepted shoulder system based on the patient's functional status to investigate various shoulder conditions led to the development of patient-related outcome rating systems. These instruments can be divided into 2 groups: global shoulder evaluation scales and pathology-focused tools (Table 18).

In 1997 L'Insalata et al.[145] published the first tested and validated global shoulder evaluation scale. They described it as "a self-administered questionnaire for the assessment of symptoms and function of the shoulder": the Shoulder Rating Questionnaire. It is unknown how the items of the instrument were selected.

"A preliminary questionnaire was developed" and "questions that had poor reliability, substantially reduced the total or subset internal consistency, or contributed little to the clinical sensitivity of the over-all instrument were eliminated to produce the final questionnaire." The final form includes 6 separately scored domains (global assessment, pain, daily activities, recreational and athletic activities, work, and satisfaction) with a series of multiple-choice questions. A weighting scheme based on "consultation with several shoulder surgeons and patients" was followed. The weighting is as follows: global assessment, 15%; pain, 40%; daily activities, 20%; recreational and athletic activities, 15%; and work, 10%. The total possible score ranges from 17 to 100. Validity and reliability were evaluated by the developers, but no a priori predictions were made and no interpretation of the observed correlations was provided. Construct validation through correlations between this instrument and other shoulder scales has not been established. However, the responsiveness for this tool has not been compared with any other existing shoulder instruments.

The American Academy of Orthopaedic Surgeons, along with the Institute for Work & Health (Toronto, Ontario, Canada), developed a 30-item checklist designed to globally evaluate "upper extremity-related symptoms and measure functional status at the level of disability."[146] This tool has good validity and reliability, and a complete user's manual is available.[147] Item generation was carried out by first reviewing the literature. Item reduction was carried out in 2 steps. Clinicians performed the initial item reduction.[148] Another criticism is the redundancy of the tool. The most attractive characteristic of this tool is that patients can complete the questionnaire before a diagnosis is established. Unfortunately, this instrument has been shown to be less responsive than other shoulder-specific instruments because it evaluates the distal upper limb, making it less efficient as a research tool in clinical trials.[149-151]

The Oxford Shoulder Score was the first shoulder-specific patient-based outcome scale, published in 1996 by Dawson et al.[152] It was created for patients having shoulder operations other than stabilization. The Oxford Shoulder Instability Questionnaire was developed in 1999 by the same authors,[153] and it was designed for patients who had been excluded from the original questionnaire, those presenting with shoulder instability. Both are 12-item questionnaires with each item scored from 1 to 5. The final score ranges from 12 (best score) to 60 (worst score). Unfortunately, it is unknown whether these patients (investigated during the tool-construction phase) represented all types of shoulder categories and treatment experiences, all ages, and both genders. It is not stated by what method the items were selected or discarded. Otherwise, these questionnaires have been extensively tested and provide reliable, valid, and responsive information.[154]

About 10 years ago, Hollinshead et al.[155] introduced a new disease-specific quality-of-life instrument indicated for use as an outcome score in patients with rotator cuff disease. The tool was constructed and tested using a methodology similar to that described by Guyatt et al.,[156] starting from a literature search, discussion with clinicians, and "direct input from a set of patients with a full spectrum of rotator cuff disease." The instrument has 34 items with 5 domains: symptoms and physical complaints (16 items), sport/recreation (4 items), work-related concerns (4 items), lifestyle issues (5 items), and social and emotional issues (5 items).[136] The authors chose a 100-mm visual analog scale response format (where 100 mm is the best score and 0 mm is the worst score). They also recommend converting the raw scores (0 to 3,400 [where 0 is the worst score and 3,400 is the best score]) to a percentage score, presenting scores out of 100. Validity and reliability of the instrument were evaluated, but its responsiveness has not been reported.

Kirkley et al.[149-151] published the most advanced series of disease-specific quality-of-life measurement tools for the shoulder. They used the methodology described by Kirshner and Guyatt.[157] Testing reliability, validity, responsiveness, and the minimally important difference for each were evaluated carefully.

The WOSI,[150] released in 1998, is for use as the primary outcome measure in clinical trials evaluating treatments for patients with shoulder instability. In 2001 the Western Ontario Osteoarthritis of the Shoulder Index was published[151]; the instrument is intended for use as the primary outcome measure in clinical trials evaluating patients with symptomatic primary OA of the shoulder. In 2003 the Western Ontario Rotator Cuff Index was proposed as the primary outcome measure in clinical trials evaluating treatments for patients with degeneration of the rotator cuff.[149] Item generation was carried out in 3 steps for all 3 of the tools, which included a review of the literature and existing instruments, interviews with clinician experts, and interviews with patients representing the full spectrum of patient characteristics. Item reduction was carried out by use of the frequency importance product (impact) from a survey of 100 patients representing the full spectrum of patient characteristics and a

correlation matrix to eliminate redundant items. The response format selected for the instrument was a 10-cm visual analog scale anchored verbally at each end. The items were assigned equal weight based on the uniformly high impact scores. Each instrument includes instructions to the patient, a supplement with an explanation of each item, and detailed instructions for the clinician on scoring. The authors recommend using the total score for the primary outcome in clinical trials but also recommend reporting individual domain scores. The scores can be presented in their raw form or converted to a percent score. Validity has been assessed through construct validation by making a priori predictions of how the instrument would correlate with other measures of health status. Responsiveness was evaluated by use of change scores after an intervention of known effectiveness.[137]

KNEE RATING SCALES

Knee rating scales can be classified by a few different factors. The first is the individual who produces the responses. Some are clinician-based, that is, the clinician produces the responses used to calculate the measurement. However, more numerous are the patient-reported outcome measures. These measures often prove more valid than clinician-based measures, because they can target the patients' complaints more directly.[158-162] Patient satisfaction has been shown to correlate most closely with outcome scores that are based on patients' subjective reporting of symptoms and function.[163]

Some knee rating scales are adapted to different kinds of patients than others (Table 19). There are scales that cater to athletic patients with ligamentous knee injuries, for example, and those that cater to patients with degenerative knee diseases such as OA.[132] Yet another distinction can be made between outcome scales and activity scales. Given the patient variability just described, studies should include both.[164] Athletic patients, for example, might have different expectations, and subject their knees to different levels of stress, than patients with OA. Activity scales make it possible to adjust for these differences, which can affect patients' reporting of symptoms and function. Patient activity level is an important prognostic variable that is not always directly related to symptoms and function.[132]

The first portion of this section will address 2 rating scales that are partly clinician-based, both of which focus on athletic patients. Discussion of patient-reported rating scales follows, eventually examining 2 scales that cater to patients with OA. The section will end with a brief review of 2 activity scales. Pertinent information will be collected in tabular form.

Clinician-Based Outcome Scales

The Cincinnati Knee Rating System combines clinician-based evaluation with patient-reported symptoms and function to arrive at a comprehensive and rigorous measure. Patients usually score lower on the Cincinnati scale than on the Lysholm scale, for example.[165,166] In its current form, the system is composed of 6 subscales that add up to 100 points: 20 for symptoms, 15 for daily and sports functional activities, 25 for physical examination, 20 for knee stability testing, 10 for radiographic findings, and 10 for functional testing.[167] The Cincinnati Knee Rating System is most often used to evaluate ACL injuries and reconstruction but has proven reliable, valid, and re-

TABLE 19. *Knee Rating Scales*

Clinician-based*	Cincinnati	Ligament injury and progress after reconstruction, HTO, meniscus repair, allograft transplant
	IKDC	Knee in general
Patient-reported	Lysholm	Progress after ligament surgery; also used to evaluate other knee conditions
	SANE	Knee in general
	KOOS	Sports injury
	ACL quality of life	Chronic ACL deficiency
	WOMAC	Osteoarthritis of the lower extremities
	Oxford	Osteoarthritis of the knee, progress after total knee arthroplasty
Activity scales	Tegner	Knee activity level based on sport or type of work
	Marx	Knee activity level based on functional element

Reprinted with permission.[178]

Abbreviations: HTO, high tibial osteotomy; IKDC, International Knee Documentation Committee; SANE, Single Assessment Numeric Evaluation.

*In conjunction with patient-reported components.

sponsive to clinical change in other disorders as well.[161,168]

The International Knee Documentation Committee has developed 2 rating scales, 1 "objective" and 1 "subjective."[169] The first is clinician-based and grades patients as normal, nearly normal, abnormal, or severely abnormal with regard to a variety of parameters that include effusion, motion, ligament laxity, crepitus, harvest-site pathology, radiographic findings, and 1-leg hop test. The final patient grade is determined by the lowest grade in any given group. The subjective rating scale asks patients to respond to questions inquiring about symptoms, sports activities, and ability to function, including climbing stairs, squatting, running, and jumping. It has been shown to be reliable, valid, and responsive when applied to a range of knee conditions, including injuries to the ligaments, meniscus, and articular cartilage, as well as OA and patellofemoral knee pain.[170,171]

Patient-Reported Outcome Scales

The modified Lysholm scale is a patient-reported measure designed to evaluate outcomes after knee ligament surgery.[172] It consists of an 8-item questionnaire and is scaled to a maximum score of 100 points. Knee stability accounts for 25 points, pain for 25, locking for 15, swelling and stair climbing for 10 each, and limp, use of a support, and squatting for 5 each.[173] Originally developed in 1982 and modified in 1985, and one of the first outcome measures to rely on patient-reported symptoms and function, the Lysholm scale has been used extensively in clinical research.[174,175] Although it has shown adequate test-retest reliability and good construct validity,[132] it has endured criticism that its reliability, validity, and responsiveness are greatest when applied to evaluation of ACL reconstruction outcomes, being less robust when applied to other knee conditions.[176,177] Because scores on the Lysholm scale have been shown to vary depending on the extent to which patients self-limit their activities, it is probably most useful in conjunction with 1 or more of the activity scales to be discussed later.[166,178]

Perhaps the simplest knee rating scale, the Single Assessment Numeric Evaluation, was designed with a specific kind of patient in mind: college-aged patients who had undergone ACL reconstruction.[179] The Single Assessment Numeric Evaluation consists of just 1 question, asking patients how they would rate their knee on a scale of 0 to 100, with 100 representing normal. Although this scale can be administered quite easily and correlates well with the Lysholm scale, it is only known to be useful with a homogeneous cohort, consisting of patients who would interpret the single broad question similarly.[132,179]

The Knee Injury and Osteoarthritis Outcome Score (KOOS) is another patient-reported measure. It consists of 5 separate scores: 9 questions for pain, 7 questions for symptoms, 17 questions for activities of daily living, 5 questions for sports and recreational function, and 4 items for knee-related quality of life.[180] It includes the 24 questions of the WOMAC, to be discussed later, and the WOMAC score can be calculated from the KOOS score.[132] The KOOS has been used to evaluate ACL reconstruction, meniscectomy, tibial osteotomy, and post-traumatic OA, and it has been validated in multiple languages.[178,181-183] It is a versatile instrument whose reliability, validity, and responsiveness have been shown in a cohort of 21 ACL reconstruction patients.[180] The subscales dealing with knee-related quality of life have been shown to be the most sensitive, and these could potentially be applied successfully to yet more knee conditions.[178]

The quality-of-life outcome measure for chronic ACL deficiency was developed with input from ACL-deficient patients and primary care sports medicine physicians, orthopaedic surgeons, athletic therapists, and physical therapists.[132] It consists of 31 visual analog questions relating to 5 categories: symptoms and physical complaints, work-related concerns, recreational activities and sports participation, lifestyle, and social and emotional health status relating to the knee.[132] The scale is specifically applicable to patients with ACL deficiency and has proven valid and responsive for this population.[184]

Whereas the rating scales discussed up until this point have been designed primarily for active or athletic patients, the rating scales that will follow are designed for patients with degenerative knee disorders. They are often used to evaluate patients who have undergone total knee arthroplasty.[132]

The WOMAC is the most commonly used rating scale for patients with knee OA.[185] It consists of 24 questions divided into 3 categories: 5 questions dealing with pain, 2 with stiffness, and 17 with difficulty performing the activities of daily living. The WOMAC has been shown to be reliable, valid, and responsive and is therefore used extensively.[185,186] Because it is focused on older patients primarily, the aforementioned KOOS scale was developed to cater to younger, more active patients.[180]

The Oxford Knee Scale is notable for its extensive incorporation of patient input into its development.[187]

The questionnaire consists of 12 multiple-choice questions, each with 5 possible responses. Testing in a prospective group of 117 patients undergoing total knee arthroplasty has shown it to be reliable, valid, and responsive.[187]

Activity Scales

The beginning of this section discussed the importance of activity scales to complement outcome rating scales, allowing investigators to adjust for differences among patients in the demand placed on the knee and expectations for recovery. The following are 2 of these activity scales.

The Tegner activity level scale aims to place a patient's activity level somewhere on a 0-to-10 scale, based on the specific type of work or particular sport performed.[173] The problem is inherent in its use of specific activities to determine activity level rather than the functional elements ostensibly necessary to perform a given activity.[178] This limits generalizability, because a specific sport or kind of work can involve different functional elements in different cultures or settings.[178] Furthermore, the Tegner scale has not been validated, although it remains widely used.[164]

The Marx activity level scale is a brief activity assessment, reported by the patient, designed to be used in conjunction with outcome measures. Its questions are function specific, rather than sport specific, and also ask for the frequency with which the patient performs the function.[164] The scale consists of 4 questions, evaluating running, cutting, decelerating, and pivoting. Patients are asked to score frequency on a 0-to-4 scale for each element, for a possible 16 total points. The Marx scale has been shown to be reliable and valid, and it is quick and easy to use.[164]

Conclusions

There is a variety of reliable, valid, and responsive knee rating scales available. The challenging choice regarding which to use will depend on the specific knee condition in question. It can be said, however, that both a general health outcomes measure like the SF-36 and an activity level scale should be used in conjunction with any of the anatomy- or disease-specific rating scales discussed.

ANKLE RATING SCALES

Outcome research regarding the ankle joint, similar to any other joint, is an important tool to evaluate the efficacy of treatment after ankle injuries. Several scoring systems for evaluating ankle injuries and treatments are commonly used.[188,189] These scoring systems provide important information about the injured patient and increase the understanding of the complexity of success or failure in terms of treatment of ankle injuries. Any scoring system should include the critical items that make the scoring system accurate, reliable, and reproducible.

An increasing number of scoring scales now exist for the evaluation of ankle injuries. In addition, different pathologies often need specific outcome scales for more accurate and valid assessment. Junge et al.[190] reported that lateral ankle sprain is the most common injury in sports medicine. Moreover, other injuries such as osteochondral defects, arthritis, and tendinopathy are also related to the ankle joint.

The most commonly used ankle scales are presented and correlated with their specific pathology (Table 20).

TABLE 20. *Ankle Rating Scales*

	Instability	Osteochondral Defect/Osteoarthritis	Tendinopathy	Global Evaluation
Clinician-based outcome scales	Good (1975) Sefton (1979) Karlsson (1991) Kaikkonen (1994) AOFAS (1994)	AOFAS (1994)	AOFAS (1994)	AOFAS (1994)
Patient-related outcome scales	AJFAT (1999) FAOS (2001) FADI (2005) FAAM (2005)	FAOS (2001) FADI (2005) FAAM (2005)	FAOS (2001) FADI (2005) FAAM (2005)	FAOS (2001) FADI (2005) FAAM (2005)

NOTE. Scales are listed in increasing order of validity and reliability.
Abbreviations: AJFAT, Ankle Joint Functional Assessment Tool; FAOS, Foot and Ankle Outcome Score.

Clinician-Based Outcome Scales

The first outcomes scale for assessment of ankle injuries was described by Good et al.[191] in 1975 to report the outcome after a reconstruction of the lateral ligaments of the ankle. They graded the outcomes as excellent, good, fair, or poor. Sefton et al.[192] in 1979 reported the outcomes after surgical reconstruction of the anterior talofibular ligament. They reported grades 1 to 4 for outcome assessment. Grade 1 is the best outcome, with full activity, including strenuous sport, and no pain, swelling, or giving way. Grade 4 is the worst outcome, with recurrent instability and giving way in normal activities, with episodes of pain and swelling.[192] The scale described by Sefton et al. was based on that of Good et al. with minor modifications.

In 1982 St Pierre et al.[193] described a new scoring system for clinical assessment after reconstruction of the lateral ankle ligaments. This scoring system is based on a separate evaluation of activity level, pain, swelling, and functional instability. Each item was judged as excellent (0), good (1), fair (2), or failure (3). The scores are summed, and the assessment is graded as excellent (0), good (1), fair (2 to 6), or failure (>6).[193]

Karlsson and Peterson[194] in 1991 published a scoring system based on 8 functional categories: pain, swelling, subjective instability, stiffness, stair climbing, running, work activities, and use of external support. Each item was allocated a certain number of points, with a total of 100 points. The scoring scale describes functional estimation of ankle function.[194]

Kaikkonen et al.[195] in 1994 evaluated 11 different functional ankle tests, questionnaire answers, and results of clinical ankle examination and created a test protocol consisting of 3 simple questions that describe the functional assessment of the injured ankle, 2 clinical measurements (range of motion in dorsiflexion and laxity of the ankle joint), 1 ankle test measuring functional stability (walking down a staircase), 2 tests measuring muscle strength (rising on heels and toes), and 1 test measuring balance (balancing on a square beam). Each test could significantly differentiate between healthy controls and patients. The final total score correlated significantly with the isokinetic strength testing of the ankle, patient-related opinion about the recovery, and functional assessment. In exact numbers, after all scores are summed up, the grade is considered excellent (85 to 100), good (70 to 80), fair (55 to 65), or poor (≤50). This scoring system is recommended for studies that evaluate functional recovery after ankle injuries.[195]

Moreover, in 1994 the American Orthopaedic Foot and Ankle Society (AOFAS) developed clinical rating scales to establish standard guidelines for the assessment of foot and ankle surgery.[196] The AOFAS clinical rating system consists of 4 rating scales that correspond to the anatomic regions of the foot and ankle: ankle-hindfoot scale, midfoot scale, hallux metatarsophalangeal–interphalangeal scale, and lesser metatarsophalangeal–interphalangeal scale. The AOFAS scoring system is the most used foot and ankle scale. The AOFAS ankle-hindfoot scoring system is based on 3 items: pain (40 points), function (50 points), and alignment (10 points). The functional assessment is divided into 7 topics: activities limitation, maximum walking distance, walking surfaces, gait abnormality, sagittal motion (flexion plus extension), hindfoot motion, and ankle instability.[196] The AOFAS rating scale has been used not only to assess ankle instability but also for other pathologies such as osteochondral defect of the talus, ankle arthritis, and tendinopathy.

In 1997 de Bie et al.[197] published a scoring system for the judgment of nonsurgical treatment after acute ankle sprain. The system is based on functional evaluation of pain, stability, weight bearing, swelling, and walking pattern, with a maximum score of 100 points. The system is used to assess the prognosis after acute injuries. It has shown good correlation with the 2- and 4-week outcomes in 81% to 97% of patients.[197]

Patient-Related Outcome Scales

The importance of the patient's perspective is becoming more and more recognized in health care and is the most important criterion for judgment of treatment outcomes.[198] Patient-assessed measures provide a feasible and appropriate method to address the concerns of the patient, for instance, in the context of clinical trials.[199]

In 1999 Rozzi et al.[200] described the Ankle Joint Functional Assessment Tool, which contains 5 impairments items (pain, stiffness, stability, strength, and "rolling over"), 4 activity-related items (walking on uneven ground, cutting when running, jogging, and descending stairs), and 1 overall quality item. Each item has 5 answer options. The best total score of the Ankle Joint Functional Assessment Tool is 40 points, and the worst possible score is 0 points.

In 2001 Roos et al.[201] described the Foot and Ankle Outcome Score. The Foot and Ankle Outcome Score is a 42-item questionnaire that assesses patient-relevant outcomes in 5 subscales (pain, other symptoms,

activities of daily living, sport and recreation function, and foot- and ankle-related quality of life). The subscale "pain" contains 9 items, the subscale "other symptoms" contains 7 items, the subscale "activities of daily living" contains 17 items, the subscale "sport and recreation function" contains 5 items, and the subscale "foot- and ankle-related quality of life" contains 4 items. Each question can be scored on a 5-point scale (from 0 to 4), and each of the 5 subscale scores is then transformed to a 0-to-100, worst-to-best score.[201] This score meets all set criteria of validity and reliability and has been judged to be useful for the evaluation of patient-relevant outcomes related to ankle ligament injuries. It also can be used to assess outcomes in patients with talar osteochondral defects, OA, and tendinopathy.

In 2005 Hale and Hertel[202] described the FADI. It is a 34-item questionnaire divided into 2 subscales: the FADI and the FADI Sport. The FADI includes 4 pain-related items and 22 activity-related items. The FADI Sport contains 8 activity-related items. Each question can be scored on a 5-point scale (from 0 to 4). The FADI and the FADI Sport are scored separately. The FADI has a total score of 104 points and the FADI Sport, 32 points. The scores of the FADI and FADI Sport are then transformed into percentages.[202]

In 2005 Martin et al.[203] described the Functional Ankle Ability Measure (FAAM). It is identical to the FADI except that the "sleeping" item and the 4 "pain-related" items of the FADI are deleted. The activities-of-daily-living subscale of the FAAM (previously called the Foot and Ankle Disability Index) now includes 21 activity-related items; the sports subscale of the FAAM remains exactly the same as the FADI Sport subscale (8 activity-related items). The rating system of the FAAM is identical to the FADI. The lowest potential score of the activities-of-daily-living subscale of the FAAM is 0 points, and the highest is 84 points. The lowest potential score of the sports subscale of the FAAM is 0 points, and the highest is 32 points.[203]

According to a systematic review of patient-assessed instruments, the FADI and the FAAM can be considered the most appropriate patient-assessed tools to quantify functional disabilities in patients with chronic ankle instability.[204]

CONCLUSIONS

Researchers planning clinical trials should select a modern instrument (developed with accurate patient input for item generation and reduction, with established validity and reliability) appropriate for the investigated condition/pathology. The most responsive instrument available should be used to minimize the sample size for the proposed study.

Stefano Zaffagnini, M.D.
Brian W. Boyle, B.A.
Mario Ferretti, M.D., Ph.D.
Giulio Maria Marcheggiani Muccioli, M.D.
Robert G. Marx, M.D., M.Sc., F.R.C.S.C.

SECTION 12

Key Statistical Principles: Statistical Power in Clinical Research

What is statistical power? Statistical power from the perspective of clinical research is the ability to detect a difference in treatment effects if one exists. It is largely a theoretical concept, but one with practical implications. This applies to any study design in which you are testing a hypothesis and can compare either treatments in 2 different groups of patients or different time points (before/after treatment) in the same patients.

Imagine a study of 2 alternative types of pain medications in which there are just a few patients available for study (Table 21, study example 1). Perhaps their medical condition is uncommon in the community where the research is taking place. We randomize patients to receive treatment A or treatment B. This randomization works, and we find that the pretreatment pain levels are equivalent between the 2 groups of patients. Both groups rate their pain as 8.5 out of a possible 10 points, with 10 being the worst pain imaginable. For the purposes of this example, all standard deviations are similar, although this is not always the case.

Table 21. *Underpowered and Overpowered Study Examples*

	Study Example 1		Study Example 2	
	Treatment A	Treatment B	Treatment C	Treatment D
Sample size	5	5	5,000	5,000
Pretreatment pain (± SD)	8.5 ± 2.3	8.5 ± 2.2	8.3 ± 1.2	8.3 ± 1.0
Post-treatment pain (± SD)	2.2 ± 2.0	6.4 ± 2.1	5.2 ± 1.3	4.9 ± 1.2
P value	.22		.01	
Power	17%		98%	

After treatment, the patients' pain levels are measured again. This time we find that the patients who received treatment A have a pain level of just 2.2 whereas those who received treatment B have a pain level of 6.4. Treatment A seems to be more effective in controlling pain in these patients, right? Not so fast. First, we must perform a statistical test to determine whether the difference between treatment groups is statistically significant. To our surprise, the test result's *P* value comes back a nonsignificant .22.

Now imagine another study in which we compare 2 other pain medications in a large number of patients (Table 21, study example 2). Perhaps their medical condition is very common. Again, we randomize the patients to treatment group—this time treatment C and treatment D. The randomization again works, and we find that patients receiving each treatment had similar pretreatment pain levels of 8.3. After treatment, we find that both groups have responded to treatment. Patients receiving treatment C now have a score of 5.2, and patients receiving treatment D now have a score of 4.9. Treatment D has a lower score, but the difference is clinically irrelevant. This time the statistical test results in a *P* value of .01, which is "statistically significant" using the usual critical *P* value criterion of $< .05$. Yet there is only a slight difference between the group means.

These findings are a function of statistical power. A study with a very small sample size may show a difference in outcomes between 2 treatment options, but a statistical test of that difference may be insignificant. Alternatively, a study with a very large sample size may find a statistically significant difference in the outcomes between 2 treatments, but the difference may be clinically irrelevant. In the first case, the study is underpowered. In the second case, it is overpowered.

WHY DOES STATISTICAL POWER MATTER?

Statistical power provides both investigators and reviewers with a sense of the ability of a study to answer the research question. Although it can be argued that no clinical study can demonstrate causation, these studies can provide guidance for treatment options and be quite valuable in improving patient care. If a study is known to be underpowered, the investigators know they must be cautious in interpreting nonsignificant results. Likewise, a reader of the study should consider the power when determining whether the results reflect "the truth" or are simply a reflection of an inadequate sample size.

Overpowering a study may be a waste of resources, time, and energy, but it may also provide the investigators with an opportunity to explore the hypothesis of interest within subgroups of patients. For example, a treatment may be found to have a very small effect in the overall study population (as found in study example 2), but perhaps on subgroup analysis, we find that women have a clinically impressive response to one treatment compared with the other but men do not. In a study overpowered to study the overall hypothesis that one treatment has better outcomes than another, there may be adequate power to identify these subgroup differences, which may be missed in a study that is only powered to detect the main association of interest. Investigators should also be cautious about over-interpreting a statistically significant effect when the effect size is small and potentially clinically irrelevant.

Underpowering a study, however, may result in missing a true treatment effect simply because a sufficient number of patients were not included in the study. This will result in a null finding when there is a true effect. In this case we miss an opportunity to improve our understanding of clinical care, and our

patients are worse off as a result. Clearly, underpowering a study is the worst of these 2 scenarios.

An increasing number of orthopaedic journals are requiring that sample size calculations be provided in submitted manuscripts to allow the reviewers and subsequent readers the opportunity to evaluate the usefulness of the study findings. An underpowered study may still be publishable, but the importance of the findings despite the lack of power will weigh much heavier in the decision to publish.

WHEN DO WE NEED STATISTICAL POWER?

Statistical power is required anytime you want to test differences—by this, we mean anytime you want to determine whether there is a statistically significant difference between groups or a statistically significant relation between 2 variables. When testing hypotheses, statistical power determines your ability to detect a difference if one truly exists.

The rationale for this is both scientific and philosophical. When conducting scientific research, the research is only worth undertaking if there is the possibility of rejecting the null hypothesis. Without adequate power, this is questionable. Underpowered research is less likely to be published or to contribute to our body of knowledge. As such, it is considered unethical (the philosophical argument) to perform underpowered research. You are subjecting humans to unnecessary inconvenience at the very least. At worst, you are subjecting humans to unnecessary interventions and, therefore, risk of harm.

If you are not formally testing hypotheses, then statistical power is not strictly necessary. However, if you are looking for correlations between 2 variables, then you still need adequate sample size (i.e., enough power). For example, if you were evaluating whether ultrasound could diagnose a rotator cuff tear as well as a more expensive MRI scan, then you would want a sample size large enough to provide you with a reliable estimate of the ability of ultrasound to correctly diagnose a rotator cuff tear. A study of 3 patients evaluated with both MRI and ultrasound would likely be inadequate to answer this question, because there are only 4 possible results: 0% accuracy (0 of 3 ultrasounds agree with the MRI), 33% accuracy (1 of 3 agree), 67% accuracy (2 of 3 agree), and 100% accuracy (3 of 3 agree). Clearly, to obtain a reliable estimate, more samples would be needed. An entire body of literature has been developed evaluating the sample size requirements of reliability studies, but such specifics lie beyond the scope of this chapter.

WHAT ARE THE PROPERTIES OF STATISTICAL POWER?

Statistical power is usually presented either as a percentage between 0% and 100% or, less commonly, as a proportion between 0.00 and 1.00. Power is calculated as $1 - \beta$, where β is a type II error, or the likelihood of rejecting the alternative hypothesis if it is true. So 0.80 power would be interpreted as having 80% power to detect a difference if it truly exists. For most clinical research, 80% power is considered the lowest acceptable figure because you have just a 1-in-5 chance of missing a true difference. For some studies, 90% power may be preferable if the consequences of missing a meaningful difference are serious.

For example, the established treatment (treatment E) is effective and relatively inexpensive but has a high risk of complications. A new experimental treatment (treatment F) is believed to be both effective and safe, but it costs substantially more than treatment E. In comparing these 2 treatments head to head, we would not want to miss a true treatment effect difference if one existed, so we might consider powering our study to more than 80%. If we missed a true treatment effect difference in treatment E's favor, it is possible that treatment E would be abandoned for treatment F even though it is more effective, simply because the study did not shown an effect difference and treatment decisions might be made based on cost alone. Conversely, if we missed a true treatment effect difference in treatment F's favor, it is possible that treatment F would not be accepted into general clinical practice because of the prohibitive costs.

Power is influenced by sample size, variability/frequency, *P* value, and effect size. Adjusting any of these characteristics changes the statistical power. Sample size is what most of us think of first when thinking about statistical power. The higher the sample size, the higher the power if all other factors remain equal. Likewise, a lower sample size will always have a lower power, all other factors being equal. This is also the most easily modifiable factor in calculation of power. We can usually recruit more patients, but it is much more difficult to justify adjusting the other components of power.

Variability is a measure of how much spread exists in the data. A measure that is highly variable between individual subjects will result in a larger standard

deviation or variance (section 14). The larger this variation, the greater the number of subjects needed will be, because any difference between the group means may be masked by the variability. This variability only applies to power calculations for continuous or scale parameters because there is no measure of variability for discrete variables.

Frequency is the alternative to variability for discrete measures. A study's power is optimized when the frequency of either a discrete dependent (outcome) or independent (explanatory) variable is balanced. A study using a variable with a much lower frequency will require many more patients to achieve adequate statistical power than a study in which the frequency is balanced across groups. For example, if 50% of the study subjects had valgus knees and 50% had varus knees, an analysis comparing knee deformities would have optimum power. If a third group of knees with no varus or valgus were included, then the optimal power would be achieved if each group represented 33.3% of the sample.

A critical *P* value of .05 is usually accepted for most clinical research. If a smaller *P* value is desired, power will be decreased, because it will be more difficult to achieve a smaller *P* value than a *P* value of .05 and a true difference may be missed. Conversely, if a larger *P* value were considered statistically significant, power would be increased. *P* values are not usually modified unless, as before with an adjustment for power, there were a reason to be more or less inclusive of what is considered a statistically significant result.

Often, *P* values will be adjusted for multiple comparisons if many different analyses are being conducted on the same subjects. By way of example, one such adjustment is known as a Bonferroni correction, in which the critical *P* value of .05 is divided by the number of comparisons being made. If there were 5 hypotheses being tested, the new critical *P* value would effectively become .01 (.05 ÷ 5 comparisons). The power would then be calculated based on this new effective *P* value.

Effect size refers to the size of the effect you expect to find or the minimally clinically relevant difference. If you do not have an expected effect size based on previous information (e.g., pilot data or other research findings from the literature), then using the minimally clinically relevant difference is most appropriate. As a rule of thumb, the minimally clinically relevant difference would be the smallest change expected to make a difference. This difference may be in a subject's health, quality of life, or satisfaction or in a myriad of other measures considered clinically important.

In orthopaedics especially, this is often scale data, such as a patient-reported outcome measure (e.g., KOOS). In the case of such scale data, the minimal difference would be the smallest difference for which a subject can actually discern a difference in his or her state of health. Usually, this is much smaller than a surgeon may expect from a treatment thought to be effective. If true, this will result in an overpowering of the study, but it may also allow for subgroup analyses to determine in which patients the treatment is most effective (or ineffective). Many such patient-reported outcome measures have established the minimally clinically relevant difference either in the initial validation study or in some early clinical study using the instrument. Finding these values in the literature will ease effect-size decisions when calculating power.

Adjusting the sample size, variability/frequency, critical *P* value, or effect size will change the power for a given study. Because most scientific journals require a *P* value of .05 or less to be considered statistically significant, this is the power characteristic least easily modifiable for a power calculation.

Variability and frequency are only really adjustable in the design of a research project. Variability can be reduced if the patient population selected for study is more homogeneous, but this will reduce generalizability of the results. Likewise, patients could be recruited based on discrete characteristics, so the frequency of these characteristics could be manipulated to achieve maximum power (e.g., recruiting 50% varus and 50% valgus knees rather than enrolling consecutive patients).

Effect sizes are also not easily amenable to adjustment, because a justifiable effect size is needed to adequately power a study. If we were to choose an unreasonably large effect size, we would be left with a lot of statistical power, but we would be very unlikely to achieve an effect size that large, so we would still have a negative result—and an underpowered study for an effect size we consider clinically meaningful.

As mentioned before, adjusting sample size is the most easily manipulable power characteristic, which is why we often equate sample size with power. If we set our *P* value, estimate our effect size, and estimate our frequency or variability, we will be left with sample size required to achieve 80% (or greater) power. Because we can usually recruit more patients, this is the simplest way to achieve adequate power. In the rare instance when more patients are not available,

modifications of the other characteristics may be required, although this is not recommended.

A special case would be a study in which we have a limited number of cases but an unlimited number of control patients available. Perhaps we want to study the factors associated with patients having pulmonary embolism (PE) after knee arthroscopy. We can only identify a limited number of patients who have had a PE after knee arthroscopy, but we can identify many, many patients who did not have a PE after knee arthroscopy. In this example, we would include all patients with a PE and then sample control patients who did not have a PE. We can manipulate our statistical power in this case by recruiting multiple controls per case. Most such case-control studies are conducted with a 1:1 control-case ratio, but power can be increased by using 2:1, 3:1, or even 4:1. The power gain becomes minimal after a 6:1 ratio, so it is not particularly useful to use more controls than that.

HOW CAN WE BE SURE OF OUR STATISTICAL POWER?

We cannot be sure of our statistical power. Statistical power is an estimate of the likelihood of finding a true difference if one exists, but it is only as accurate as the estimates that we provide. If our estimates of effect size are overly generous, we may be underpowered for the actual effect size found. If our variability is higher than anticipated, we will lose power. Only our *P* value and sample size are more reliable, but even for sample size, it is not uncommon for patients to drop out of a study before completing follow-up; thus, if an adequate number of patients are not recruited to make up for these losses, the study will lose power.

Ideally, all research projects should have an a priori power calculation. In some cases investigators fail to calculate power a priori, in which case they should certainly calculate post hoc power to inform themselves and others about the value of the study findings. Even for studies in which an a priori power calculation was performed, it is sometimes useful to calculate power post hoc if there are appreciable differences between the estimates provided a priori and the actual results of the study.

HOW DO YOU CALCULATE STATISTICAL POWER?

Very few statisticians calculate power by hand any longer. Most use statistical software programs to calculate statistical power. Both stand-alone programs and macros for common statistical packages, such as SAS (SAS, Cary, NC) or S-Plus (TIBCO, Palo Alto, CA), are available. Stata (StataCorp LP, College Station, TX) also has some built-in power calculations available.

Web-based power calculators have proven unreliable—and are for use at your own risk because you do not know whether the underlying calculation is coded properly. This kind of mistake is much less likely with professional statistical software packages designed to calculate power.

For a surgeon interested in performing clinical research, the most appropriate way to determine your needed sample size and potential statistical power is by consulting with a statistician. If you do not have a statistician available for consultation, it is worthwhile to invest in a sample size program. Several free programs are available online (REFS), although PASS (Power Analysis and Sample Size; NCSS, Kaysville, UT) is currently the most powerful sample size calculator available, with calculations available for more than 150 statistical tests. If the analytic plans for your research, which should also be determined based on a consult with a statistician, tend to be relatively basic (e.g., statistics described in section 13), a free program may be sufficient for your sample size calculation needs. PASS may be overkill in those circumstances. If you are unable to use these free programs, then an online calculator is the source of last resort.

CONCLUSIONS

Statistical power is an often misunderstood and sometimes abused theoretical concept with practical implications. Conducting an underpowered study is a waste of time and is potentially a violation of a physician's responsibility to first do no harm. An overpowered study is less troubling but still may waste resources and time that could have been devoted to other efforts.

Power is influenced by sample size, variability/frequency, *P* value, and effect size. Adjusting any of these characteristics changes the statistical power, although in most circumstances sample size is the most easily changeable characteristic. Ideally, power should be calculated a priori (before starting the study), although a post hoc power calculation may also be useful if study characteristics are very different from what was estimated before beginning the research. Fortunately, calculating statistical power is relatively

easy with today's modern statistical software programs, many of which are available free of charge.

SUGGESTED READING

Jones SR, Carley S, Harrison M. An introduction to power and sample size estimation. *Emerg Med J* 2003;20:453-458.

Adcock CJ. Sample size determination: A review. *Statistician* 1997;46:261-283.

Sim J, Wright CC. The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Phys Ther* 2005;85:257-268.

Saito Y, Sozu T, Hamada H, Yoshimura I. Effective number of subjects and number of raters for inter-rater reliability studies. *Stat Med* 2006;25:1547-1560.

Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. *Stat Med* 1998;17:101-110.

Carley S, Dosman S, Jones SR, Harrison M. Simple nomograms to calculate sample size in diagnostic studies. *Emerg Med J* 2005;22:180-181.

Shiboski S. UCSF Department of Epidemiology & Biostatistics. Power and Sample Size Programs. Available from: http://www.epibiostat.ucsf.edu/biostat/sampsize.html.

Stephen Lyman, Ph.D.

SECTION 13

Key Statistical Principles: Common Statistical Tests

How does the investigator determine whether the differences observed in a study are truly significant? Subjective clinical experience may be necessary to determine clinical significance, but statistical significance can be calculated with statistical tests. In this section, we discuss several commonly used statistical tests and present examples of the types of research questions that each is designed to help answer. The relevant parameters that determine which test is most appropriate for analyzing a given data set are explained, and the equations that are used for each type of test are presented. Specifically, we discuss the following tests: *t* tests, Mann-Whitney *U* test, χ^2 and Fisher exact tests, analysis of variance (ANOVA), Kruskal-Wallis test, and Generalized Estimating Equations (GEE). The flowchart shown in Fig 10 represents the outline of this section and provides a graphic comparison of the assumptions underlying each of these tests. Using this flowchart, the investigator can quickly determine which test is most appropriate for his or her data set when the dimension (how many groups are being compared), distribution (whether or not data points are normally distributed), and dependency (whether the variables are dependent or independent) of the data are known.

TWO-SAMPLE PARAMETRIC TESTS

A parametric test is built on a specific distribution, and by convention, it assumes a normally distributed population in practice. In this section we focus on the most popular parametric test, the *t* test, for either 2 independent or dependent populations (matched pairs or repeatedly measured samples).

t Tests

General Assumptions of *t* Tests

Theoretically, *t* tests can be used when the sample sizes are very small (e.g., <30) and the primary assumptions for *t* tests include the following:

- The population distribution from which the sample data are drawn is normal.
- The populations have equal variances.

The normality assumption can be verified by looking at the distribution of the data using histograms or by performing a normality test (e.g., Kolmogorov-Smirnov test). The equality-of-variances assumption is usually examined by an *F* test using statistical software.

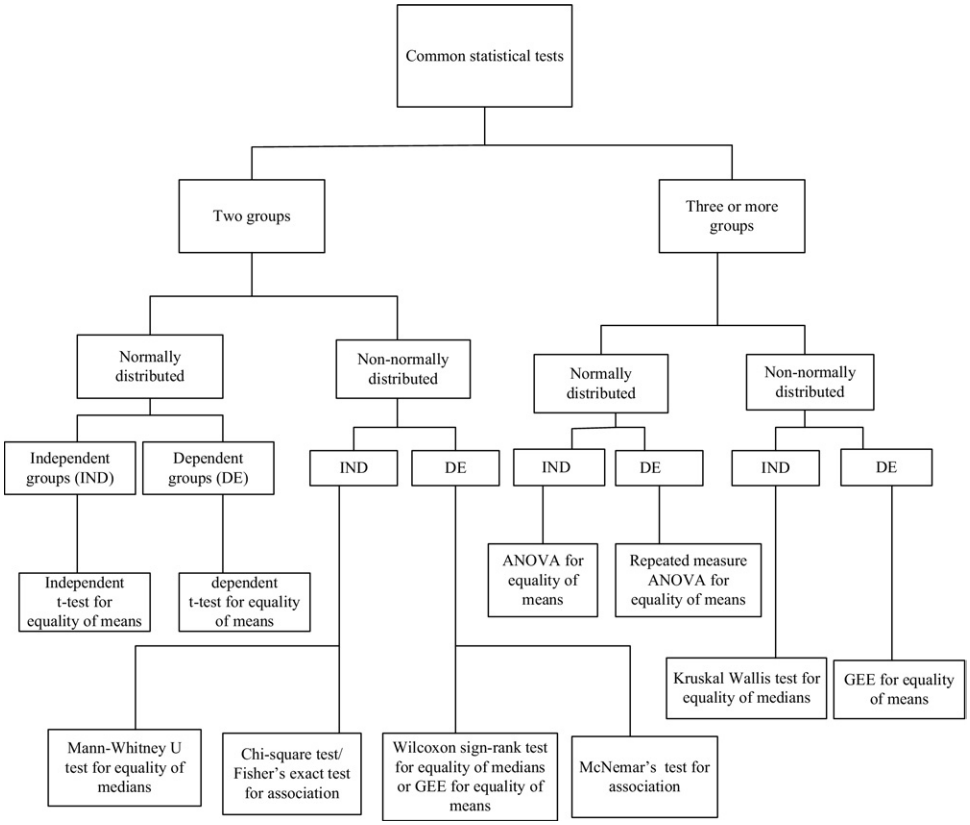

FIGURE 10. A flowchart of the commonly used statistical tests.

As an example, let us consider an RCT of patellofemoral OA treatment with one group of 20 patients receiving treatment A and another group of 20 patients receiving treatment B. One of the outcomes of interest is the knee flexion before and after treatment. Table 22 shows the sample mean and sample standard deviation of the change in knee flexion after treatment.

Examples of common hypotheses in this type of study include the following:

Hypothesis 1. H_0: The post-treatment mean knee flexion in group A = 115 versus H_1: The post-treatment mean knee flexion in group A ≠ 115.

Hypothesis 2. H_0: The mean change in knee flexion in group A = the mean change in knee flexion in group B versus H_1: The means are different.

Hypothesis 3. H_0: The mean change in knee flexion in group A = 0 versus H_1: The mean change in knee flexion in group A ≠ 0.

The three hypotheses can potentially be solved by the most frequently used *t* tests: 1-sample *t* test, independent 2-sample *t* test, and paired samples *t* test, respectively.

One-Sample *t* Test

A 1-sample *t* test is used to test whether the population mean μ is equal to a specified value μ_o with the test statistic:

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

where *t* follows a *t* distribution with $(n - 1)$ degrees of freedom under the null hypothesis of $\mu = \mu_o$ and $\bar{x}$ is

TABLE 22. *Change in Knee Flexion: An Example for* t *Test*

	Change in Knee Flexion (Postoperatively-Preoperatively)	
	Treatment A	Treatment B
Sample mean ($\bar{x}$)	10	5
Sample SD (s)	10	9

the sample mean, s is the sample standard deviation, and n is the sample size.

Example 1: If the sample mean (standard deviation) of post-treatment knee flexion in group A is 110 (10), and the specified "standard" value is 115, the test statistic for testing hypothesis 1 (above) is as follows:

$$t = \frac{110 - 115}{\frac{10}{\sqrt{20}}}$$

The P value is computed in statistical software by comparing the test statistic t with the critical value $t_0 = t_{(0.025,(n-1))}$. In this case the P value is less than .05, indicating that the mean post-treatment knee flexion in group A is significantly different from 115.

Independent 2-Sample *t Test*

The independent 2-sample t test is used to test whether the means of 2 independent populations are equal under the null hypothesis. Different formulae have been developed for the following scenarios.

Equal Sample Sizes, Equal Variance: This test can be used when the 2 samples have the same number of subjects ($n_1 = n_2 = n$) and the 2 distributions have the same variance with the test statistic:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{2}{n}}}$$

where $s_p = \sqrt{\frac{s_1^2+s_2^2}{2}}$ is the pooled standard deviation and the denominator of t represents the standard error of the difference between the 2 means. The test statistic t follows a t distribution with $2(n - 1)$ degrees of freedom.

Example 2: Under the assumption of equal variance, the test statistic for hypothesis 2 (above) is:

$$t = \frac{10 - 5}{s_p\sqrt{\frac{2}{20}}}$$

where $s_p = \sqrt{\frac{10^2+9^2}{2}}$ and the degrees of freedom is $2 \times (20 - 1)$.

The P value is greater than .05, implying that there is no significant difference in change of knee flexion between the 2 groups.

Unequal Sample Sizes, Equal Variance: When the 2 samples have a different number of subjects ($n_1 \neq n_2$) but the 2 distributions have the same variance, the test statistic is:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where $s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$. The test statistic follows a t distribution with $(n_1 + n_2 - 2)$ degrees of freedom.

Equal Sample Sizes, Unequal Variance: When the 2 sample sizes are the same ($n_1 = n_2 = n$) but the variances are assumed to be different, the test statistic is:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2 + s_2^2}{n}}},$$

following a t distribution with $\frac{(n-1)(s_1^2+s_2^2)^2}{s_1^4+s_2^4}$ degrees of freedom.

Example 3: Under the assumption of unequal variance, the test statistic for hypothesis 2 (above) is:

$$t = \frac{10 - 5}{\sqrt{\frac{10^2 + 9^2}{20}}}$$

with $\frac{(20-10)(10^2+9^2)^2}{10^4+9^4}$ degrees of freedom. The P value is greater than .05, implying that there is no significant difference in change of knee flexion between the 2 groups.

Unequal Sample Sizes, Unequal Variance: When the 2 sample sizes are different ($n_1 \neq n_2$) and the variances are assumed to be different, the test statistic is:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

following a t distribution with $\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}$ degrees of freedom.

Dependent 2-Sample *t Test*

When the same sample is measured twice or 2 samples are matched, the dependent 2-sample t test

can be used to test the difference between the means of the dependent outcomes. The test statistic is:

$$t = \frac{\bar{x}_d - \Delta}{\frac{S_d}{\sqrt{n}}}$$

following a t distribution under the null hypothesis with $(n - 1)$ degrees of freedom, where $\bar{x}_d$ denotes the mean of the difference, S_d is the standard deviation of the difference, and Δ is the hypothetic difference in the null hypothesis.

Example 4: Because the knee flexion was measured on the same sample (patient) before and after treatment, the dependency between the repeated measurements should be taken into account when testing hypothesis 3. The test statistic is:

$$t = \frac{10 - 0}{\frac{10}{\sqrt{20}}}$$

with (20 – 1) degrees of freedom. The P value is less than .05, implying that the mean knee flexion after treatment is significantly different from the mean knee flexion before treatment.

TWO-SAMPLE NONPARAMETRIC TESTS

A nonparametric test is distribution free, meaning data are not assumed to come from any specific distributions. In practice, as an alternative to parametric tests, nonparametric tests are applied in particular when sample size is small or data are not normally distributed.

The Mann-Whitney U test and χ^2/Fisher exact test are used when the variables are independent, whereas the Wilcoxon signed-rank test and McNemar test are used when variables are dependent.

Mann-Whitney *U* test

The Mann-Whitney U test is a nonparametric test for assessing whether 2 independent groups are equally distributed. The test can be applied to ordinal or continuous data without assuming normality. It is an alternative to the independent 2-sample t test, when the assumption of normality is not met. It would be used to test hypothesis 2 (above) if the samples in groups A and B were equally distributed. Assume the 2 groups A and B have sample sizes n_A and n_B, respectively. To apply the Mann-Whitney U test, raw data from the entire sample combining groups A and B are ranked from smallest to largest, with the smallest value receiving a rank of 1. Ties are assigned average ranks. The test statistic U is a function of these ranks:

$$U = n_A n_B + \frac{n_A(n_A + 1)}{2} - R_A$$

where R_A denotes the sum of ranks for group A.

Wilcoxon Signed-Rank Test

The Wilcoxon signed-rank test is a nonparametric analog to the paired t test, and it can be used when the differences between pairs are not normally distributed. The test is often conducted to assess the difference between values of outcome data before and after an intervention with hypothesis H_0: the median difference = 0 versus H_a: the median difference ≠ 0.

Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ represent n paired samples and $D_i = X_i - Y_i$; $i = 1,2, \ldots, n$, the difference between pairs. The absolute values of D_i are ranked from smallest to largest and the test statistic $W = \min(W_+, W_-)$ is a function of the ranks R_i, where $W_+ = \sum_{i=1}^{n} I(D_i > 0)R_i$, $W_- = \sum_{i=1}^{n} I(D_i < 0)R_i$ are the sums of the ranks for positive differences and negative differences, respectively.

Example 5: Shown in Table 23 are details for the calculation of the Wilcoxon signed-rank test statistic for SF-12 mental health composite scores from a sample of 10 patients before and after total knee replacement.

Note that in the case of a tie, the mean of the ranks is taken. For example, subjects 5 and 10 have the same value of $|X_1 - X_2|$. The mean of their ranks is $\frac{5+6}{2} = 5.5$.

The sum of ranks with a positive sign in $X_1 - X_2$ is $W_+ = 9 + 1 + 7 = 17$, and that of ranks with a negative sign is $W_- = 8 + 2 + 5.5 + 4 + 3 + 10 + 5.5 = 38$. Hence the test statistic $W = \min(17, 38) = 17$ with $P = .3$, indicating that there is no significant difference between SF-12 scores before and after total knee replacement.

χ^2 Test

Contingency tables are commonly used in clinical research to describe the relation between the row and the column variables. In these types of analyses, 2 groups with independent variables are compared.

For example, Table 24 is a 2 × 2 contingency table of the incidence of nausea in patients receiving either

TABLE 23. *Mental Health Composite Scores in SF-12: An Example for Wilcoxon Signed-Rank Test*

Subject	Before Surgery Score: X_1	After Surgery Score: X_2	Sign of $X_1 - X_2$	$\lvert X_1 - X_2 \rvert$	Rank of $\lvert X_1 - X_2 \rvert$
1	80	65	+	15	9
2	60	72	−	12	8
3	55	62	−	7	2
4	70	66	+	4	1
5	85	95	−	10	5.5
6	83	92	−	9	4
7	66	74	−	8	3
8	52	92	−	40	10
9	73	62	+	11	7
10	80	90	−	10	5.5

general anesthesia or regional anesthesia during total knee replacement.

Let $o_{1,1}$, $o_{1,2}$, $o_{2,1}$, and $o_{2,2}$ denote the observed frequency of each combination of anesthesia type and incidence of nausea in cells $c_{1,1}$, $c_{1,2}$ $c_{2,1}$, and $c_{2,2}$, and $e_{1,1}$, $e_{1,2}$, $e_{2,1}$, and $e_{2,2}$ denote the corresponding expected frequency if there were no association, where the expected frequency:

$$e_{i,j} = \frac{row\ i\ total \times column\ j\ total}{total\ (N)},\ i, j = 1, 2$$

One way to assess the association between the row and the column variables is by measuring the difference between the observed and expected frequencies with a χ^2 test. The χ^2 test statistic is defined by the following:

$$\hat{\chi}^2 = \sum_{(i,j)} \frac{(o_{i,j} - e_{i,j})^2}{e_{i,j}}$$

The test can be applied to any row-by-column, or $m \times n$, contingency table, $m,n > 1$. Under the null hypothesis of no association between the row and the column variables, the test statistic follows a χ^2 distribution with $(m - 1)(n - 1)$ degrees of freedom.

Example 6: The expected frequencies in Table 24 are as follows:

TABLE 24. *Association Between Anesthesia Type and Incidence of Nausea*

Anesthesia Type	Incidence of Nausea: Yes	Incidence of Nausea: No	Row Total
General	$O_{1,1} = 70$ ($c_{1,1}$)	$O_{1,2} = 20$ ($c_{1,2}$)	90
Regional	$O_{2,1} = 30$ ($c_{2,1}$)	$O_{2,2} = 60$ ($c_{2,2}$)	90
Column total	100	80	Total N = 180

$$e_{1,1} = \frac{90 \times 100}{180} = 50,\ e_{1,2} = \frac{90 \times 80}{180} = 40,$$

$$e_{2,1} = \frac{90 \times 100}{180} = 50,\ e_{2,2} = \frac{90 \times 80}{180} = 40$$

Hence,

$$\hat{\chi}^2 = \frac{(70 - 50)^2}{50} + \frac{(20 - 40)^2}{40} + \frac{(30 - 50)^2}{50} + \frac{(60 - 40)^2}{40} = 36$$

and the degree of freedom is $(2 - 1) \times (2 - 1) = 1$. The P value obtained with statistical software is less than .05, implying that patients who received general anesthesia during total knee replacement are more likely to have nausea than patients who received regional anesthesia.

Fisher Exact Test

When the expected frequency in any cell of a contingency table is less than 5, the χ^2 test becomes inaccurate and loses its power because it relies on large samples. For example, in a study comparing mortality rates between patients undergoing unilateral or bilateral knee replacement, the incidence of death is very low, resulting in highly unbalanced data allocations among the cells of the table. In such a case, the Fisher exact test is an alternative to the χ^2 test. Because, in general, the computation of the Fisher exact test is not feasible by hand, we avoid the detailed formula here.

McNemar test

If the χ^2 (or Fisher exact) test could be considered the independent 2-sample t test for categorical variables, the McNemar test is the counterpart of the

TABLE 25. *Case-Control Study of Cancer*

	Non-cancer patient		
	Smoker	Non-smoker	Row Total
Cancer patient			
Smoker	a	b	a + b
Non-smoker	c	d	c + d
Column total	a + c	b + d	Total N

paired *t* test for comparing dependent categorical variables. For example, the investigators studied the association between smoking and lung cancer in a case-control study where *N* cancer patients (cases) were matched with *N* non-cancer patients (controls) in Table 25 based on age, gender, location, and other related variables. In this case the χ^2 test and the Fisher exact test are not appropriate because they assume that the samples are independent.

The McNemar test is a modification of the χ^2 test, taking into account the correlation between the matched samples. Because the concordance cells where both case and control are smokers (a) or non-smokers (d) do not provide information about the association between cancer and smoking, the McNemar test only contains the frequencies in the discordance cells (b and c) and is defined as:

$$\hat{\chi}^2 = \frac{(|b - c| - 0.5)^2}{b + c}$$

The test statistic follows a χ^2 distribution with 1 degree of freedom under the null hypothesis of no association between cancer and smoking.

MULTIPLE-SAMPLE PARAMETRIC TESTS

These tests are used when comparing data sets among 3 or more groups. The ANOVA is used for normally distributed samples/data, whereas the Kruskal-Wallis test and GEE are appropriate for samples with normal or non-normal distributions.

One-Way ANOVA

A 1-way ANOVA is an alternative to the independent 2-sample *t* test for testing the equality of 3 or more means by use of variances.

The assumptions of ANOVA include the following:

- The samples are drawn from populations following normal distributions.
- The samples are independent.
- The populations have equal variances.

The null hypothesis of ANOVA is that all population means are equal, and the alternative hypothesis is that at least one population mean is different.

The basis of ANOVA is to partition the total variation into "between-group variation" and "within-group variation" and compare the two. These and other terms related to ANOVA are defined below.

Grand mean is the average of all sample values.

Between-group variation is the sum of squared differences between each group mean and the grand mean. The between-group variance is the between-group variation divided by its degrees of freedom. If there are *g* groups, the degrees of freedom is then equal to $g - 1$.

Within-group variation is the sum of squared differences between each sample and its group mean. The within-group variance is the within-group variation divided by its degrees of freedom. If there are *g* groups and *n* samples within each group, the degrees of freedom is then equal to $g(n - 1)$ or $N - 1$, where *N* is the total sample size.

Total variation is the sum of between-group variation and within-group variation.

The ANOVA is used to compare the ratio (*F* test statistic) of between-group variance to within-group variance. If the between-group variance is much larger

TABLE 26. *One-Way ANOVA*

Source	Sum of Squares (Variation)	Degrees of Freedom	Mean Square (Variance)	*F* Statistic
Between group	SSB	$g - 1$	$\text{MSB} = \frac{SSB}{g-1}$	$\frac{MSB}{MSW}$
Within group	SSW	$g\,(n - 1)$	$\text{MSW} = \frac{SSW}{g(n-1)}$	
Total	SST = SSB + SSW	$N - 1$		

Abbreviations: SSB, sum of squares between groups; SSW, sum of squares within groups; SST, total sum of squares; MSB, mean squares between groups; MSW, mean squares within groups; g, number of groups; n, number of samples within each group.

TABLE 27. *Two-Way ANOVA*

Source	Sum of Squares	Degrees of Freedom	Mean Square	*F* Statistic
Main effect 1	SS1	$g_1 - 1$	MS1 = SS1/*df*	MS1/MSW
Main effect 2	SS2	$g_2 - 1$	MS2 = SS2/*df*	MS2/MSW
Interaction effect	SS12	$(g_1 - 1)(g_2 - 1)$	MS12 = SS12/*df*	MS12/MSW
Within	SSW	$g_1 g_2 (n - 1)$	MSW = SSW/*df*	
Total	SST = SS1 + SS2 + SS12 + SSW	$g_1 g_2 n - 1$		

Abbreviations: SS1, sum of squares for Main Effect 1; SS2, sum of squares for Main Effect 2; SS12, sum of squares for interaction between Main Effect 1 and Main Effect 2; SSW, sum of squares within groups; SST, total sum of squares; MS1, mean squares for Main Effect 1; MS2, mean squares for Main Effect 2; MS12, mean squares for interaction between Main Effect 1 and Main Effect 2; MSW, mean squares within groups; g, number of groups; n, number of samples within each group.

than the within-group variance, then we conclude that the means are different. This is summarized in an ANOVA table (Table 26).

Two-Way ANOVA

In contrast to 1-way ANOVA, which tests the equality of population means in one variable, 2-way ANOVA is extended to assess the difference among population means in 2 independent variables or factors.

The 2-way ANOVA has the same assumptions as the 1-way ANOVA.

The null hypotheses in a 2-way ANOVA include:

- Main effect: The population means of each factor are equal.
- Interaction effect: There is no interaction between the 2 factors.

Similar to 1-way ANOVA, 2-way ANOVA partitions the total variation into 2 main effects or between-group variations, within-group variation, and interaction effects between the 2 factors. There is an *F* test for testing each main effect and the interaction effect. A similar table is created for 2-way ANOVA (Table 27).

MULTIPLE-SAMPLE NONPARAMETRIC TESTS

Kruskal-Wallis Test

The Kruskal-Wallis test is a generalization of the Mann-Whitney *U* test for testing the equality of 3 or more population medians and is a nonparametric alternative to 1-way ANOVA. Like other nonparametric tests, the Kruskal-Wallis test is based on the ranks of data and does not assume normality.

Assume there are g independent groups with n_i observations in the i group, $i = 1, 2, \ldots, n$. To calculate the Kruskal-Wallis test statistic, rank all data from the g groups with the smallest value obtaining a rank of 1. Ties are assigned average ranks. The test statistic is given by the following:

$$K = (n-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_i - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2}$$

where r_{ij} is the rank among all data of observation j in group i, $\bar{r}_i$ is the mean rank of all observations in group i, and $\bar{r}$ is the mean rank of all observations across all groups.

The test statistic K follows a χ^2 distribution under the null hypothesis with $g - 1$ degrees of freedom. The P value can be obtained from the χ^2 distribution table.

Correlation Analysis

In addition to the previous statistical tests, we next briefly discuss correlation analysis. A variety of correlation coefficients are available and used to assess the relation between 2 or more random variables. We introduce 2 commonly used correlation coefficients, the Pearson correlation and Spearman rank correlation coefficients.

Pearson Correlation: Pearson correlation, also called Pearson product-moment correlation, was developed by Karl Pearson. It is applied to continuous variables and assumes a linear relation between 2 normally distributed variables. Pearson correlation lies in [–1, 1], with 1 (–1) indicating a perfect positive (negative) linear relationship. For a pair of independent variables, the Pearson correlation is 0.

Spearman Rank Correlation: Spearman rank correlation is a nonparametric correlation. When 2 variables are not normally distributed or do not have a linear relation, Spearman rank correlation is an alternative to Pearson correlation. Like those nonparametric tests we introduced earlier, Spearman rank correlation is also calculated based on ranks and therefore is not affected by the distribution of data.

ADVANCED STATISTICAL TESTS

In addition to those commonly used statistical tests, there are advanced statistical methods available for more complicated data settings. A couple of examples are presented here.

Regression Analysis

Regression analysis is a method for assessing the relation between a dependent variable and one or more independent variables. The most commonly used regression analysis is linear regression, which assumes a linear relation between the dependent and independent variables.

Repeated-Measures ANOVA

As with any ANOVA, repeated-measures ANOVA tests the equality of multiple means. However, repeated-measures ANOVA is used when the same group of random samples is measured under the same condition at multiple time points or under different conditions. It assumes data to be normally distributed and can be considered an extension of the paired *t* test to a sample with more than 2 repeated measures.

The GEE Method

The GEE method is for modeling clustered data and longitudinal data. When data are clustered dependent, the GEE allows for fitting the parameters of a generalized linear model without explicitly defining the correlation structure.

CONCLUSIONS

Statistical tests prove that observed differences are not due to random chance, providing scientific rigor to clinical and other experimental findings. Examples in this section show that specific tests have been developed to analyze most types of data sets that are of interest to the academic clinician-scientist. As outlined in this section, the appropriate test for a given data set is simple to determine based on 3 basic aspects of the data set(s): dimension (whether 2 or more groups are being compared), distribution (whether data are normally or non-normally distributed), and dependency (whether variables are dependent or independent). In the context of clinically relevant study design and interpretation of results, statistical tests establish nonrandom correlations that rigorously support efficacy, safety, or other outcomes of therapeutic interventions or other factors that are of interest to the clinician-scientist investigator.

SUGGESTED READING

Agresti A. *Categorical data analysis.* Hoboken, NJ: Wiley, 2002.

Rumsey D. *Statistical II for dummies.* Hoboken, NJ: Wiley, 2009.

Norman GR, Streiner DL. *PDQ statistics.* Ed 3. Shelton, CT: People's Medical Publishing House, 2003.

Pagano M, Gauvreau K. *Principles of biostatistics.* Ed 2. Pacific Grove, CA: Duxbury Press, 2000.

Kahn HA, Sempos CT. *Statistical methods in epidemiology.* Oxford: Oxford University Press, 1989.

Petrie A. Statistics in orthopaedic papers. *J Bone Joint Surg Br* 2006;88:1121-1136.

Hanley JA, Negassa A, Edwardes MD, Forrester JE. Statistical analysis of correlated data using generalized estimating equations: An orientation. *Am J Epidemiol* 2003;157:364-375.

Yan Ma, Ph.D.
Chisa Hidaka, M.D.
Stephen Lyman, Ph.D.

SECTION 14

Key Statistical Principles: The Nature of Data

The goal of clinical research is to use information collected from a sample of patients to answer questions about all patients with that condition or who receive that treatment. The statistical inference methods we use to do this requires that (1) the selected sample is representative of the population of interest and (2) we know something about the distribution of the data. If a variable's distribution approximates that

of a known probability distribution, like the normal distribution, that has well-described parameters, then we can use our knowledge of these distributions to calculate the probability that our hypotheses are valid using parametric tests (described in section 13). When a variable does not follow such a distribution, we need to use different methods that do not rely on parameters to help us, using nonparametric tests (also described in section 13). This makes understanding our data important in developing the proper analysis plan for our study, and the point of this chapter is to:

1. describe the tools used to summarize data and learn about the distributions (descriptive statistics),
2. develop methods for estimating association between two variables (measures of Association),
3. describe how we use our knowledge of the parameters of probability distributions to confirm or negate our hypotheses (confidence intervals and *P* values).

DESCRIPTIVE STATISTICS

Every variable has an underlying distribution function that describes how the observations are spread over all possible values of the variable. This distribution is influenced by the type of variable: categorical (discrete) or continuous. In brief, continuous variables are those that can take on any value in a range of possible values, whereas categorical variables can only take on a specific set of values. Categorical variables can be further classified as ordinal, where the categories have a defined order (e.g., disagree, neutral, agree), or nominal, where there is no intrinsic order (e.g., gender [male, female]). The purpose of descriptive statistics is to generate a few measures that give us an idea of the particular features of the distribution of the variable of interest.

FREQUENCIES AND PERCENTAGES

A simple way to describe the distribution of a variable is to list all of the different values and the frequency of each value (i.e., the number of times the value occurs in the data set) (Table 28).

We can see that presenting frequencies and percentages in a table is an effective way to describe categorical data because there are a limited number of possible values. This method does not work as well for numerical variables because the range of possible values is often much larger and it is not practical to display all the observed values. One way to resolve this problem, if it makes sense for the analysis, is to group the values of the variable into a smaller number of defined categories and calculate the frequencies and percentages for these created categories. This is often done with variables such as age (e.g., ≤59 years, 60 to 79 years, and ≥80 years) and clinical laboratory values such as serum vitamin D level (where <20 ng/mL is deficient, 20 to 31 ng/mL is insufficient, and ≥32 ng/mL is sufficient), where groupings of the numeric data make sense clinically. If it does not make sense to categorize the variable, other methods are necessary to summarize the data.

NUMERICAL SUMMARY METHODS

Measures of Location

These are methods to describe the center of a distribution of continuous data. The mean and median are most common, although the mode is rarely used in special circumstances.

Mean: The mean of a variable is the sum of all values of the variable divided by the number of observations. In statistical notation this is represented by the following:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

So, using the variable age from our example data in Table 29, the mean age is as follows:

$$\text{Mean age} = (49 + 45 + 63 + 41 + 50 + 29 + 37 + 40 + 40 + 47)/10 = 441/10 = 44.1$$

The mean is simple to compute and has nice theoretical properties in terms of statistics that make it widely used. However, the mean is sensitive to extreme values, especially when the number of observations is small.

Median: The median of a variable is the central value when the data points are arranged in rank order, so that half of the data values are higher than the

TABLE 28. *Sex Distribution in Hypothetical Study Population*

	N	%
Sex		
Female	437	53.2
Male	384	46.8
Total	821	100.0

TABLE 29. *Age for 10 Hypothetical Research Subjects*

Subject	Age (yr)	Squared Deviation From Mean $(x_i - \bar{x})^2$
1	49	24.01
2	45	0.81
3	63	357.21
4	41	9.61
5	50	34.81
6	29	228.01
7	37	50.41
8	40	16.81
9	40	16.81
10	47	8.41
	Mean ($\bar{x}$), 44.1	Sum, 746.9

median value and the other half are lower than the median value. When there are an even number of observations in a data set, the median is defined as the midpoint between the 2 middle values. In our age example, the median is calculated as follows:

Age sorted lowest to highest : 29, 37, 40, 40, 41, 45, 47, 49, 50, 63

This data set has an even number of observations, so the 2 middle values are 41 and 45. The median is (41 + 45)/2 = 43.

The median requires the data to be sorted, so it is not as simple to compute as the mean, especially with larger sample sizes. It is not as sensitive to outlying values, though, so it may be a better measure of central tendency, especially for smaller samples.

Mode: The mode of a variable is the value that occurs most frequently. A multimodal variable has more than 1 value that meets this criterion. In our age example the mode is 40, because it occurs twice and all other values occur only once. This statistic is not often reported because it is usually not useful for describing continuous variables, which may have all unique values so that every value in the data set is a mode.

Measures of Variability/Dispersion

Range: The range is the difference between the largest and smallest values of a variable (Table 29). A wider range indicates more variability in the data. In our age data, the range is 63 – 29 = 34. The minimum and maximum values of a variable are more often reported than the range, however, because these 2 values also provide some information about the location of the extremes of a variable.

Variance: The variance of a data set, denoted s^2, is a measure of variability around the sample mean. The equation for variance is listed below. In words, the variance is the average of the squared deviations from the mean:

$$\text{Variance}(s^2) = \frac{1}{(n-1)}\sum_{i-1}^{n}(x_i - \bar{x})^2$$

$$= \frac{1}{(20-1)}\sum_{i-1}^{n}(x_i - \bar{x})^2$$

$$\text{Variance} = \frac{1}{(9)}(764.9) = 83.0$$

Standard Deviation: The standard deviation is the square root of the variance. This value is more often reported in descriptive tables because it is measured in the same units as the variable, and the mean and standard deviation together can tell us a lot about a distribution of values.

MEASURES OF ASSOCIATION

The objective of the study designs discussed in previous chapters is to compare outcomes in 2 or more groups, such as new treatment versus old treatment, exposure versus no exposure, and so on. The numerical summary methods we have discussed above are useful in describing individual variables, but now we need to define some measures of association that will let us compare groups.

Relative Risk (Risk Ratio)

What Is Relative Risk? In epidemiology, the incidence of an event (e.g., disease diagnosis or surgical complication) is the frequency of new events that occur during a specified time interval. What we call the "risk" of an event is the incidence rate, which is the incidence divided by the population at risk during that time interval.

$$\text{Incidence rate} = \frac{\text{Number of new events}}{\text{Population at risk for event}}$$

A related concept is the prevalence of the event, which is the sum of events that have already occurred plus new incident events divided by the population total. So we can see that the incidence rate is a measure of the risk of disease whereas prevalence shows the burden of disease in a population.

As formulas, incidence and prevalence can be described as follows:

$$\text{Incidence} = \text{Number of knee replacement surgeries performed this year}$$

TABLE 30. *Two-by-Two Frequency Table*

	Outcome		
	Yes	No	
Treatment A	a	b	a + b
Treatment B	c	d	c + d
	a + c	b + d	

$$\text{Incidence rate} = \frac{\text{Number of Knee arthroplasties performed this year}}{\text{Number of people in population}}$$

$$\text{Prevalence rate} = \frac{\text{Number of people living with knee arthroplasty}}{\text{Number of people in population}}$$

What we usually want to do in clinical research is to compare the risk between groups, so an easy way to do this is to simply determine the ratio of the risk (whether incidence or prevalence) for the 2 groups:

$$\text{Relative risk} = \frac{\text{Risk in group 1}}{\text{Risk in group 2}}$$

How to Calculate the Relative Risk in a Cohort Study: When preparing data for a comparison of 2 treatments, we can most easily calculate relative risk by creating a 2 × 2 table (Table 30). The term "2 × 2" refers to the numbers of rows and columns in the table—2 possible outcomes and 2 possible treatments in the example in Table 30.

We can calculate the likelihood (or risk) of the outcome for each treatment group as follows:

$$\text{Risk}_{\text{TXA}} = \frac{A}{a + b}$$

$$\text{Risk}_{\text{TXA}} = \frac{C}{c + d}$$

We can then calculate a relative risk of the outcome in patients receiving treatment A versus treatment B:

$$\text{Relative risk} = \frac{\text{Risk}_{\text{TXA}}}{\text{Risk}_{\text{TXB}}} = \frac{a/(a + b)}{c/(c + d)}$$

How to Interpret the Relative Risk: If the relative risk equals 1, then the risk is the same in both groups and there does not appear to be an association. When the relative risk is greater than 1, the risk in group 1 is greater than that in group 2; this is usually described as evidence of an increased risk, or positive association. If the relative risk is less than 1, the risk in group 1 is less than that in group 2; this is usually described as indicating a negative association, or a decreased risk.

For example, if we had a cohort study that wanted to determine whether a new form of pain management reduced the incidence of postoperative pain, we would generate a contingency table from our data (Table 31). The relative risk is less than 1, which indicates that the new pain management technique reduces the incidence of postoperative pain.

Odds Ratio (Relative Odds)

Calculating relative risk requires us to know the incidence rate for a population, which is not possible for some study designs. In a case-control study, for example, the 2 groups are based on outcome status, so we do not know the population at risk. Thus we need another measure of association that will work for both cohort and case-control studies. For this type of study, we can calculate a different measure of association, using the odds.

What Is an Odds Ratio? The odds of an event is defined as the ratio of the probability that an event occurs to the probability that the event does not occur. If we represent the probability that event A occurs by P, then the probability that event A does not occur is 1 – P. So the odds of event A is as follows:

$$\text{Odds} = \frac{P}{1 - P}$$

For example, when rolling a die, the probability of rolling a 1 or 2 is 2/6 = 1/3 = 33.3%, so the odds of rolling a 1 or 2 is as follows:

$$\text{Odds} = \frac{33.30\%}{66.70\%} = 0.50$$

It is important to note that the probability of rolling a 1 or 2 (33.3%) and the odds of rolling a 1 or 2 (0.50) are 2 distinct measures.

How to Calculate the Odds Ratio: Now suppose we have a study as in the contingency table for a

TABLE 31. *Sample Data for Relative Risk*

	Postoperative Pain	No Postoperative Pain	Total
New pain management	4	21	25
Old pain management	9	16	25
Total	14	36	

NOTE. Risk in patients with new technique = 4/25 = 0.16. Risk in patients with old technique = 9/25 = 0.36. Relative risk = 0.16/0.36 = 0.44.

cohort study as shown in Table 32. In a cohort study we are comparing the odds of event A in the exposed group with the odds of event A in the non-exposed group.

First, we need to calculate the probability (P) of event A for group 1:

$$P = a/(a + b)$$

Next, we will calculate the odds of event A for group 1:

$$\text{Odds} = [a/(a + b)]/[b/(a + b)] = a/b$$

Similarly, the odds of event A for group 2 equals c/d. Finally, the odds ratio for group 1 versus group 2 is (a/b)/(c/d) = ad/bc (Table 33).

In a case-control study, first, we need to calculate the odds that a case had a history of exposure ($\text{Odds}_{\text{cases}}$):

$$\text{Odds}_{\text{cases}} = [a/(a + c)]/[c/(a + c)] = a/c$$

Next we calculate the odds that a control had a history of exposure ($\text{Odds}_{\text{controls}}$):

$$\text{Odds}_{\text{controls}} = [b/(b + d)]/[c/(b + d)] = b/d$$

$$\text{Odds ratio} = \text{Odds}_{\text{cases}}/\text{Odds}_{\text{controls}} = (a/c)/(b/d) = ad/bc$$

$$\text{Odds ratio} = \frac{(a/c)}{(b/d)} = \frac{ad}{bc}$$

Note that the formula for the odds ratio is the same for both cohort and case-control studies.

How to Interpret the Odds Ratio: Similar to the interpretation of the relative risk, an odds ratio of 1 indicates that the exposure is not related to the event.

If the odds ratio is larger than 1, then the exposure is positively associated with the event, and if the odds ratio is less than 1, the exposure is negatively associated with the event.

Using the Odds Ratio to Estimate Relative Risk: The odds ratio is itself a useful measure of association, but there may be situations when reporting the relative risk is preferred. In a case-control study, although the relative risk cannot be directly calculated, the odds ratio is a good approximation of the relative risk when the cases and controls are representative samples of the populations from which they are drawn and the outcome is infrequent.

We will use examples of a cohort study, where both the relative risk and odds ratio can be directly calculated, to see when the odds ratio is a good estimate of the relative risk.

When event is infrequent:

When event is frequent:

TABLE 32. *Contingency Table for a Cohort Study*

	Event A		
	Yes	No	
Exposed	a	b	a + b
Not exposed	c	d	c + d
	a + c	b + d	

TABLE 33. *Odds of Event A for Group 1*

	Event A		
	Cases	Controls	
History of exposure	a	b	a + b
Not exposed	c	d	c + d
	a + c	b + d	

	Event A		
	Yes	No	
Exposed	25	975	1,000
Not exposed	10	990	1,000
	35	1,965	

$$\text{Relative risk} = \frac{25/1,000}{10/1,000} = \frac{25}{10} = 2.50$$

$$\text{Odds ratio} = \frac{25 \times 990}{10 \times 975} = \frac{24,750}{9,750} = 2.54$$

	Event A		
	Yes	No	
Exposed	250	750	1,000
Not exposed	100	900	1,000
	350	1,750	

$$\text{Relative risk} = \frac{250/1,000}{100/1,000} = \frac{250}{100} = 2.50$$

$$\text{Odds ratio} = \frac{25 \times 990}{10 \times 750} = \frac{24,250}{9,900} = 3.00$$

MEASURES OF PROBABILITY

Understanding the properties of a distribution allows us to apply this knowledge to the first steps of

understanding statistical inference, which is the process of drawing conclusions about an entire population based on the information from a sample of that population. Recall that it is our goal to describe or make an educated estimate of some characteristic of a continuous variable using the information from our sample of observations.

There are 2 ways of estimating these characteristics. "Point estimation" involves taking the sample data and calculating a single number, such as the mean, to estimate the parameter of interest. However, the inherent problem of calculating 1 mean from 1 sample of a population is that drawing a second sample and calculating its mean may yield a very different value. The point estimate does not take into account the inherent variability that exists between any combinations of samples that are drawn from all populations. To account for this variability, a second technique, called "interval estimates," provides a reasonable range of values that are intended to contain the parameter of interest with a certain degree of confidence. This range in values is called a confidence interval (CI).

The CI allows us to evaluate the precision of a point estimate by calculating an interval that contains the true population mean with a planned degree of certainty. For the 95% CI, we are 95% confident that the true population mean lies somewhere between the upper and lower limits calculated. Another way to understand the 95% CI is as follows: if we were to select 100 random samples from a population and use these samples to calculate 100 different intervals for these samples, 95 of these intervals would cover the true population mean (whereas 5 would not).

A few ways not to interpret the CI is to state that the probability of the calculated mean lies between the upper and lower limits of the calculated interval. In addition, it would be incorrect to state that there is a 95% chance that the mean is between the upper and lower limits of the calculated interval.

Suppose, for example, that we were looking to find the CI for serum cholesterol for all men in the United States who are hypertensive and smoke. If the mean serum cholesterol level in a sample of 12 hypertensive men is 217 mg/100 mL with a standard deviation of 46 mg/100 mL, what is the 95% CI for this calculated mean? To calculate the upper and lower bounds, we first use the equation for the interval for a continuous variable:

$$\bar{X} \pm 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$$

where X is the calculated mean, σ is the standard deviation, and n is the sample population. When the values are plugged into the equation, we end up with a lower limit of 191 and an upper limit of 243. Although the interval values calculated appear to indicate a fairly precise mean, what would you imagine would happen if we were able to increase the sample size of the sample we collected? Imagine that all that changed from this sample was only the number of our sample. If the sample size was increased from 12 to 50, the CI now changes to 204 in the lower limit and 230 in the upper limit. As you can see, the sample size plays an important role in the precision of our estimates. The more people we have in our study, the more narrow the range, which in turn increases our accuracy.

As stated earlier, the bounds of the CI give us an important indicator of the precision of the calculated mean. Therefore the more narrow the CI, the more precise the estimate. The CI is an important and extremely helpful way of evaluating an estimate and, if possible, should always be reported whenever an estimate is provided in the results. Whereas the standard deviation gives the reader an idea of the spread of the values around the mean, the CI provides the reader the precision of the estimate.

CONCLUSIONS

Up until now, the chapters of this book have focused on designing studies. This chapter begins to explore what to do with the data once they have been collected. The first step should be to describe each variable. For continuous variables, we calculate the appropriate measures of central tendency and spread or dispersion. For categorical variables, we create frequency tables and calculate percentages for each stratum within each variable. Next, we calculate measures of association through either a relative risk if we know the underlying distribution or an odds ratio if we have conducted a case-control study. Finally, we calculate measures of probability. This can be done through hypothesis testing as described in section 13 but also through the calculation of CIs, which gives us a different perspective on the probability underlying our data.

SUGGESTED READING

Hennekens CH, Buring JE, Mayrent SL. *Epidemiology in medicine.* Philadelphia: Lippincott Williams & Wilkins, 1987.

Rothman KJ, Greenland S, Lash TL. *Modern epidemiology.* Ed 3. Philadelphia: Lippincott Williams & Wilkins, 2008.

Hulley SB, Cummings SR, Browner WS, Grady D, Hearst N, Newman TB. *Designing clinical research.* Ed 2. Philadelphia: Lippincott Williams & Wilkins, 2001.

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.

Gordis L. *Epidemiology.* Philadelphia: Elsevier Health Sciences, 1996.

Morshed S, Tornetta P III, Bhandari M. Analysis of observational studies: a guide to understanding statistical methods. *J Bone Joint Surg Am* 2009;91: 50-60 (Suppl 3).

Petrie A. Statistics in orthopaedic papers. *J Bone Joint Surg Br* 2006;88:1121-1136.

Varkevisser C, Pathmanathan I, Brownlee A. Designing and conducting health systems research projects, volume 2. Data analyses and report writing. KIT, IDRC: 2003.

Pagano M, Gauvreau K. *Principles of biostatistics.* Pacific Grove, CA: Duxbury Press, 2000.

Huong T. Do, M.A.
Joseph Nguyen, M.P.H.
Stephen Lyman, Ph.D.

SECTION 15

Survival Analysis in Orthopaedic Surgery: A Practical Approach

Survival analysis is an effective statistical tool for evaluating and comparing outcomes of orthopaedic procedures. This method is based on constructing a life-table of a cohort of patients after certain orthopaedic procedures. The life-table contains all the data relevant for the determination of the cohort at regular follow-up periods. The main outcome value in the life-table is the cumulative survival of the study group at each time interval with provision of 95% CIs of distribution of cumulative survival values. The calculation of these values is based on the recognition of a number of patients who were lost to follow-up and determination of the uniform criteria for patients with failed outcome. If the latter parameters are similar in different studies, a comparison of survival values can be performed by the log-rank test.

To evaluate the clinical outcome of orthopaedic procedures, 2 important and unique characteristics should be addressed: the relatively limited number of patients (<100 patients in most studies) and the term of follow-up (usually several years). These requirements might challenge the effectiveness of traditional statistical tools for comparison of medical or surgical treatments used in other clinical areas, with involvement of large cohorts of patients with clear short-term outcomes that remain unchanged for long time periods. To answer this specific need, orthopaedic procedures are evaluated and compared by use of survival analysis, which has been especially adapted to the field of orthopaedic surgery. Initially, this method was developed for the long-term follow-up of prosthetic implants,[205] but it can also be used for other orthopaedic procedures.[206]

There are 2 main methods for survivorship analysis. In the classic "product limit method" according to Kaplan and Meier, the survival (i.e., the success of the procedure) changes immediately after clinical failure.[207] Using this method in relatively small groups of evaluated patients, the CIs at the change points of the survivorship might be misleadingly overestimated or even show values above 100%.[208] Therefore, for more reliable evaluation of orthopaedic procedures with relatively small groups of patients who are followed up at constant time intervals, for example, on an annual basis in the arthroplasty follow-up, a need for special adaptation of this method is apparent. Exactly for this purpose, Murray et al.[208] popularized a method of survivorship analysis based on construction of a "life-table" with the assumption that all the procedures were performed at the same time 0 and the patients

TABLE 34. *Life-Table of Patients Operated on in 1989-1994 With BioModular Uncemented Total Shoulder Prosthesis*[212]

Postoperative Year	No. at Start	Success	Lost	Died	Failed	Withdrawn at Last Review				
						No. at Risk	Proportion Failing (%)	Proportion Succeeding (%)	Cumulative Survival (%)	95% Confidence Interval
1	90	0	0	1	5	89.5	5.6	94.4	94.4	87.6-97.6
2	84	0	0	1	3	83.5	3.6	96.4	91	83.1-95.4
3	80	0	0	0	4	80	5.0	95.0	86.5	77.6-92.2
4	76	0	0	0	5	76	6.6	93.4	80.7	70.9-87.8
5	71	0	0	2	2	70	2.9	97.1	78.4	68.1-86.0
6	67	0	0	2	2	66	3.0	97.0	76.1	65.5-84.3
7	63	1	1	4	2	60	3.3	96.7	73.5	62.4-82.2
8	55	7	0	2	0	50.5	0	100	73.5	62.1-82.4
9	46	4	0	1	1	43.5	2.3	97.7	71.8	59.9-81.3
10	40	18	0	0	1	31	3.2	96.8	69.5	56.8-79.8
11	21	9	0	1	0	16	0	100	69.5	55.3-80.7

NOTE. Postoperative years 1, 7, and 8 represent the data discussed in the text.

re-evaluated at constant intervals, taking into consideration patients who were lost to follow-up, thus establishing a cumulative success rate for each time interval. Subsequently, according to these considerations, 95% CIs of survival were determined. In this method 95% CIs are more appropriate for a small group of patients and never exceed 100% of survivorship.

VARIABLES

As an example of a life-table (Table 34), we use data published on survival analysis of 90 patients after total shoulder arthroplasty.[209] According to the method presented here, the main outcome values are the cumulative survival rates for each time period with 95% CI distribution of these values. The survival values can be presented graphically as survival curves. In addition to these final outcome values, the life-table includes all the parameters that are required for the calculation of the main outcome values; thus it contains all the data for independent evaluation of survivorship outcome, enabling critical review by readers and an ability to compare outcomes with other studies. The calculation method is shown in rows 1, 7, and 8 in Table 34.

TIME PERIODS OF FOLLOW-UP

In the first column of the life-table, the follow-up periods are given. As has been noted, the main characteristic of the presented survival analysis is the constant periods between patient evaluations according to the nature of the surgical procedure. In the presented example, because the life-table deals with the outcome of shoulder arthroplasty, 1 year between follow-up evaluations is a commonly used practice. Because the purpose of the survival analysis, among others, is a comparison between different cohorts of patients, the use of the established follow-up period for the particular procedure is recommended. An additional basic assumption of this method is that all the patients were treated at time 0. This does not mean that all the patients actually underwent surgery on the same date, but the date of the surgery for each patient is considered as time 0, after which all the calculations are performed. Accordingly, in row 1 of the life-table, the first column contains the values of 1 year; in row 7, the value of 7 years; and in row 8, the value of 8 years (i.e., 1, 7, and 8 years of follow-up).

NUMBER OF PATIENTS REMAINING FOR FOLLOW-UP AT EACH PERIOD (NUMBER AT START)

The number of patients at the start represents the number of patients who were available for evaluation at each time period. This value is a product of subtraction of the number of patients who were withdrawn from the number of patients at the start in the previous time period. Note that the number at the start in the first row (i.e., in the first time period) represents the total number of patients enrolled in the study. The number of patients withdrawn for each time period is the sum of values given in columns 3, 4, 5, and 6 (success, lost, died, and failed). The method to determine these values is given in the next section. There-

fore, in our example, in year 1, the number of patients at the start was 90 (the entire cohort). In year 7, this value is 63, when the 4 patients "withdrawn at last review" (0 + 0 + 2 + 2 = 4) were subtracted from the original number; there were 67 patients in row 6. Similarly, in row 8, the original number of patients is the product of subtraction of 8 patients (1 + 1 + 4 + 2 = 8 "withdrawn at last review" in row 7) from 63 patients, which is the original number of patients in row 7, giving a value of 55 patients.

WITHDRAWN AT LAST REVIEW

This section requires special attention because it is based on assumptions that can influence the entire life-table and can be manipulated according to special characteristics of the study group. This section contains 4 subsections (4 columns)—success, lost, died, and failed—which will be discussed separately.

Success

This might be misleading terminology, but it means that the patients reached their maximal follow-up time period and should be considered for withdrawal in the discussion of the next time period of the survival analysis. For example, in row 7, 1 patient reached the maximal follow-up of 7 years; therefore he cannot be discussed as part of the group of patients in row 8. In addition, from inspection of the life-table, the "success" column indicates the minimal follow-up time in the studied group and the number of patients who did not reach the maximal follow-up period, excluding those who were lost to follow-up and died, and at what quantitative extent. By looking at our example, we see that only 9 patients reached the whole 11-year period of follow-up, as indicated in row 11, and the minimal follow-up time was 7 years, because the first "success" is indicated in row 7.

Lost

The patients who were lost to follow-up are the main factor of uncertainty of a life-table and survival analysis. The designers of this method reasonably argued that this group might have a higher proportion of unsatisfied persons with failed procedures.[210] We will address this topic in the following sections.

Died

Two factors are crucial in the estimation of this group. It must be verified at the highest possible extent that the cause of death is unrelated to the procedure for which survival analysis is performed, because in that case the patient should be included in the "failed" group. In addition, maximal effort should be exerted to verify that the persons who have died are not included in the "lost-to-follow-up" group. The reason for the latter is that the proportion of failures in patients who died might be overestimated.[210] This might affect the other parameters of the life-table, as will be discussed later.

Failed

The way these data are filled is determined by the survival analysis constructor and has the highest potential to be biased. Unfortunately, because different authors consider different criteria for determination of failure of the studied procedure, their life-tables might be difficult for meaningful comparison. The minimalistic approach for determination of failure and the most often used is eventual revision surgery. The maximalistic approach might involve clinical signs on imaging modalities, such as radiographic signs of prosthesis loosening, a certain level of pain, restricted range of movements, and so on, without surgery. These signs can also be the reason for the decision on revision surgery[209] and become part of the minimalistic approach. Therefore a clear definition of the criteria of "failure" should be provided. It is also possible to perform a survival analysis with different failure definitions on the same group of patients to compare life-tables from different sources.

NUMBER OF PATIENTS AT RISK

This variable reflects the number of patients who are actually considered for evaluation in the certain period of time, according to the life-table design. These patients were available for follow-up at a certain time period and therefore were determined as a product of subtraction of unavailable patients, meaning those who died, were lost, or reached the end of their follow-up (success), from the total number of patients at the start of this time period. These patients at risk can reach clinical failure as discussed before, and would be removed from further follow-up, or could be considered as successes and be followed up in the next time period. The fact that not all of the subtracted individuals were exposed to the risk during the total time period should be taken into consideration. It will be impossible to know the exact fraction of these patients; therefore a reasonable estimation of 50% is used, and subtraction of only half of the

withdrawn patients is implemented for the life-table. In the example in Table 34, the number at risk in row 1 was 89.5 after subtraction of 0.5 [(0 success + 0 lost + 1 died)/2 = 0.5] from 90 (number at start).

PROPORTION OF FAILING

This is a proportional value of failed cases from the number at risk. It is usually represented in percentages. In our example (Table 34), in postoperative year 7, the proportion of failing was 3.3% (2 [failed]/60 [number at risk] × 100 = 3.3%).

PROPORTION OF SUCCEEDING

Naturally, the proportion of succeeding will be the remainder value from the proportion of failing to 100%. So, during the seventh postoperative year, the proportion of succeeding is 96.7% (100% – 3.3% [proportion failing] = 96.7%).

CUMULATIVE SURVIVAL

This is the main outcome value of the life-table and can be later represented graphically as a survival estimation for the given time period.[206] Because it is cumulative in definition, this value is calculated by multiplying the proportion succeeding in the given time period by the cumulative survival proportion in the previous time period, expressed in percentages. In the first time period, the cumulative survival proportion is equal to the proportion of succeeding, because we consider the initial cumulative survival of the procedure as 100%, as expressed in the example in Table 34. Another example is the cumulative survival of 73.5% in postoperative year 8 (1 [proportion of succeeding in year 8] × 0.735 [cumulative survival in year 7] × 100 = 73.5%).

95% CONFIDENCE INTERVAL

The last column to be filled in the life-table contains the CIs of the cumulative survival and represents distribution of 95% of these values for every time period. The calculation of the CI for a given time interval is based on determination of the "effective number of risk" (M), which contains information on the number of patients at risk from the previous time intervals according to the following formula:

$$M = i/\sum 1/n_i$$

where i is the time interval and n is the number of patients at risk in the time interval i.[208,210]

Accordingly, the confidence limits (CL) are calculated according to the following formula[209,211]:

$$CL = \frac{M}{M + 1.96^2}\cdot\left[P + \frac{1.96^2}{2\cdot M} \pm 1.96\sqrt{P\cdot\frac{(1-P)}{M} + \left(\frac{1.96^2}{4\cdot M^2}\right)}\right]$$

when M is an effective number at risk and P is cumulative survival at the given time interval (expressed as proportion and not as percentage). This mathematical expression is based on the theoretical assumption presented by Rothman[212] and popularized by Murray et al.[208] The mathematical basis of these assumptions will not be discussed in this presentation, which is more of a practical nature. The interested reader is referred to these extensive statistical reports that are given in the "References" section.

As an example of the calculations of the CIs, we will refer to time interval 8 (i = 8 [postoperative year 8]) (Table 34). The M value is 69.739 according to the following calculation:

$$\frac{8}{\frac{1}{89.5}+\frac{1}{83.5}+\frac{1}{80}+\frac{1}{76}+\frac{1}{70}+\frac{1}{66}+\frac{1}{60}+\frac{1}{50.5}} = 69.739$$

The values of the CI are calculated as follows (M = 69.739, P = 0.735).

For the upper limit,

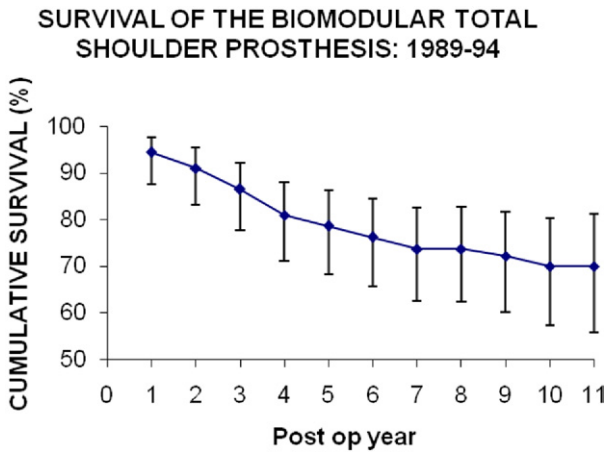

FIGURE 11. Graphic representation of outcome values of survival analysis given in Table 34. Vertical bars represent 95% confidence intervals of the cumulative survival rates.

TABLE 35. *Life-Table of Patients With Shoulder Osteoarthritis Operated on in 1989-1994 With BioModular Uncemented Total Shoulder Prosthesis*[212]

Postoperative Year	No. at Start	Success	Lost	Died	Failed	Withdrawn at Last Review: No. at Risk	Proportion Failing (%)	Proportion Succeeding (%)	Cumulative Survival (%)	95% Confidence Interval
1	48	0	0	1	4	47.5	8.4	91.6	91.6	80.3-96.7
2	43	0	0	0	3	43	7	93	85.2	72.1-92.8
3	40	0	0	0	3	40	7.5	92.5	78.8	64.9-88.2
4	37	0	0	0	3	37	8.1	91.9	72.4	57.4-83.6
5	34	0	0	0	1	34	2.9	97.9	70.9	55.5-82.7
6	33	0	0	0	2	33	6.1	93.9	66.6	50.8-79.4
7	31	0	0	3	1	29.5	3.4	96.6	64.3	48.2-77.7
8	27	3	0	1	0	25	0	100	64.3	47.7-78.1
9	23	1	1	0	1	22	4.5	95.5	61.4	44.4-76
10	20	10	0	0	0	15	0	100	61.4	43.4-76.7
11	10	2	0	1	0	8.5	0	100	61.4	41.6-78

$$\frac{M}{M+1.96^2}\cdot\left[P+\frac{1.96^2}{2\cdot M}+1.96\sqrt{P\cdot\frac{(1-P)}{M}+\frac{1.96^2}{4\cdot M^2}}\right]=0.824$$

For the lower limit,

$$\frac{M}{M+1.96^2}\cdot\left[P+\frac{1.96^2}{2\cdot M}-1.96\sqrt{P\cdot\frac{(1-P)}{M}+\frac{1.96^2}{4\cdot M^2}}\right]=0.621$$

Therefore the 95% CI for the cumulative survival of 73.5% in postoperative year 8 (Table 34) is between 62.1% and 82.4%.

At this stage, when all the data are entered into the life-table, the main outcome values, cumulative survival and its 95% CIs, can be presented graphically (Fig 11).

COMPARISON BETWEEN SURVIVAL ANALYSES

The last step of the process of evaluating the results in the life-table is the ability to compare it with the results of other survival analyses. It is clear that the prerequisite for such comparison will be the same determination for "failure" in the compared life-tables, similar numbers of "lost to follow-up," and a similar method of life-table construction.

For comparison of 2 life-tables with a relatively small number of patients with low failure rates, the log-rank test is usually used.[206] The null hypothesis of this type of comparison is the same proportion of failures in every time interval for 2 compared treatments. Using this test, we will be able to compare the occurrence of failures in the 2 survival analyses in question. For this purpose, a χ^2 statistic is calculated. For comparing 2 life-tables, the χ^2 distribution of values with 1 degree of freedom is assumed.[213] In this case the value of χ^2 above 3.841 indicates a P value below .05; when the value of χ^2 is above 6.635, the P value is below .01; and when the value of χ^2 is above 10.828, the P value is below .001.[213] We will demonstrate the calculations by using 2 life-tables (Tables 34 and 35).

For calculation of the χ^2 statistic according to the log-rank test, additional variables are determined and summarized (Table 36). "Postoperative year," "Number at risk" and "Observed failure" are taken from the life-tables that are compared.

Total number at risk" is the sum of "Number at risk" from the 2 life-tables for each postoperative year. For example, for year 7, this value is 89.5 (60 [Table 34] + 29.5 [Table 35]).

"Expected failure" for each of the life-tables for every postoperative year is calculated according to the following formula:

(Observed failure)
× (Number at risk)/(Total number at risk)

In our example, in postoperative year 7 in Table 35, this value is 0.33 (1 [observed failure] × 29.5 [number at risk]/89.5 [total number at risk]).

After the previously described variables are determined, the χ^2 statistic can be calculated for each of the life-tables according to the formula (Observed failures

TABLE 36. *Variables Required for Comparison of Survival Data in Tables 34 and 35*

Postoperative Year	No. at Risk: Table 34	Observed Failure: Table 34	No. at Risk: Table 35	Observed Failure: Table 35	Total No. at Risk	Expected Failure: Table 34	Expected Failure: Table 35
1	89.5	5	47.5	4	137	3.27	1.39
2	83.5	3	43	3	126.5	1.98	1.02
3	80	4	40	3	120	2.67	1.00
4	76	5	37	3	113	3.36	0.98
5	70	2	34	1	104	1.35	0.33
6	66	2	33	2	99	1.33	0.67
7	60	2	29.5	1	89.5	1.34	0.33
8	50.5	0	25	0	75.5	0.00	0.00
9	43.5	1	22	1	65.5	0.66	0.34
10	31	1	15	0	46	0.67	0.00
11	16	0	8.5	0	24.5	0.00	0.00

– Expected failures)^2/Expected failures, summing up to the postoperative year in question. In comparing 2 life-tables, χ^2 is equal to the sum of the results of this formula for each in the example above. If we compare the 11-year survival from Tables 34 and 35, χ^2 equals 26.07 according to the following calculation: (25.00 [sum of observed failures until year 11 in Table 34] – 16.63 [sum of expected failures until year 11 in Table 34])^2/16.63 [sum of expected failures until year 11 in Table 34] + (18.00 [sum of observed failures until year 11 in Table 35] – 6.05 [sum of expected failures until year 11 in Table 35])^2/6.05 [sum of expected failures until year 11 in Table 35]) = (25.00 – 16.63)^2/16.63 + (18.00 – 6.05)^2/6.05 = 26.07.

This value of χ^2 is higher than 10.828, giving a P value $< .001$. Therefore the difference in the 11-year survival of the implanted shoulder prostheses between these 2 groups of patients is highly significant.

CONCLUSIONS

A method for constructing and comparing survival analyses of orthopaedic procedures by use of the life-table method is presented. The method requires simple arithmetical calculations and can be further simplified by use of basic computer software, such as commonly used spreadsheet software packages. The main issue that should be addressed in this method of survival analyses is a determination of the endpoint criteria for "failures."

Nahum Rosenberg, M.D.
Michael Soudry, M.D.

SECTION 16

Outcome Measures in Multicenter Studies

Small communication errors between different project teams can result in a catastrophic failure: for example, the loss of radio contact between NASA and its Mars Climate Orbiter in 1999 led to a loss of more than US $125 million.[214] The metric/US customary unit mix-up that destroyed the craft was caused by human error in the software development and therefore severe communication problems associated with a lack of control. This nonmedical case exemplifies the need for appropriate harmonization, communication, and subsequent control if more than one group is involved in a complex research project.

Orthopaedic multicenter studies are complex by nature. They are difficult to organize, complex to manage, and hard to analyze. However, there are good reasons to face these challenges:

1. The larger sample size enables testing hypotheses with greater statistical power. It also allows

a more precise estimation of population parameters.[215] Especially in low-prevalence disorders, multicenter studies represent the sole option to generate a large enough sample size.

2. The findings and observations of multicenter studies are more generalizable than those of 1 single-center only.[216] The heterogeneity in patient demographics, clinical characteristics, and treatment differences contributes to the variance in study outcome. Even if the treatment is uniformly delivered, it may result in different outcomes at different sites (e.g., European sites compared with Asian sites).
3. The study protocol as a result of a consensus process of experts from different sites is more likely to represent the general opinion in a field and has a better chance for acceptance in the scientific community after the study.[215] This has been recently demonstrated in a large cross-sectional survey of 796 surgeons. The majority of them agreed that they would change their practice based on the results of a large randomized trial.[217]

CHALLENGES IN MULTICENTER STUDIES

The advantages of multicenter studies represent a number of challenges at the same time. The inclusion of more study sites increases the complexity. Slight differences in treatment modalities have to be considered. Working processes that may work locally without extensive infrastructure (e.g., patient monitoring) are not feasible at another site. In most studies differences in infrastructure between various sites require an independent system for data acquisition and processing. The inclusion of several study sites also requires strict monitoring to obtain a defined level of data quality. In summary, it has to be ensured that all sites measure the same variable with the same instrument and the same quality.

Although the inclusion of sites with different cultural background makes the study more representative, this is one of the greatest challenges in multicenter studies. It leads to a number of confounding variables such as socioeconomic environment or different patient expectations. Inclusion of non–English-speaking sites requires cross-cultural adaptation with translation and validation of questionnaires. Differences in cultural background have to be considered during interpretation of data.

Finally, different legal and ethical boundary conditions aggravate study preparation, performance, and analysis. Necessary applications to local ethics committees are becoming more and more complex, time-consuming, and expensive. Different legal restrictions add another challenge in multicenter studies.

The necessary infrastructure and manpower lead to increased costs and time compared with single-center studies. All these challenges have to be considered during planning and performance of multicenter studies to avoid major pitfalls and to produce valuable data.

In summary, there are 2 main challenges related to outcome measures in multicenter trials:

1. Measuring the same data. This means that at 1 site, exactly the same variable is measured as at the other site.
2. Obtaining the same data. This means that varying infrastructure as well as different legal, socioeconomic, and cultural boundary conditions may influence parameters locally, which aggravates further data processing and analysis.

This article should help to identify key components related to outcome measures in multicenter studies. Examples will be used to illustrate possible pitfalls but also strategies to avoid them.

OBJECTIVE OUTCOME MEASURES IN MULTICENTER STUDIES

Although objective outcome parameters are considered as investigator independent, there are a number of factors that may increase variability or introduce sources of unsystematic or systematic errors in multicenter studies. If parameters are measured with different devices, different protocols, or different setups, further data processing may be aggravated.

Range of Motion

Active range of motion and passive range of motion are the most widely used orthopaedic measures in daily clinical practice as well as in clinical studies. Despite their widespread use, there exists a great variability in recording methods. Whereas one group quantified standard errors of measurement between 14° and 25° (interrater trial) and between 11° and 23° (intrarater trial) when comparing 5 methods for assessing shoulder range of motion,[218] other authors concluded in a systematic review that "inter-rater reliability for measurement of passive physiological movements in lower extremity joints is generally low."[219] If objective instruments are used, the inter-

rater reliability of passive physiologic range of motion at the upper extremity can be improved.[220] For example, sufficient intrarater and interrater reliability could be demonstrated when a Fastrak measurement system (Polhemus, Colchester, VT) was used for measuring cervical spine flexion/extension, lateral flexion, and rotation and shoulder flexion/extension, abduction, and external rotation in healthy subjects.[221] However, these systems require handling know-how, are costly, and are often not available at all study sites. Therefore recording of active and passive range of motion with simple methods like the goniometer has to be standardized across study sites. This includes exact definition of measurement planes, starting position, the neutral position for the joint, and the description of the course of movement. The values should be recorded in the neutral-zero method. If only the complete range or a deficit in one plane is reported, information about changes in the neutral position are lacking.[222]

Active and passive range of motion should be recorded with the neutral-zero method. Each movement should be described exactly in the study protocol.

Performance Tests

Physical function and its impairment due to disease activity can be quantified with performance tests. Most tests include the recording of the time needed for a patient to perform the requested activity. In addition, several observational methods have been described that use ratings from observers to assess the quality of physical function.[223] However, Terwee et al.[223] found a number of methodologic shortcomings during a review of the measurement properties of all performance-based methods that have been used to measure the physical function of patients with osteoarthritis of the hip or knee. Most of the tests in this study showed low values for reliability, which represents a challenge for multicenter studies. Impellizzeri and Marcora[224] propose that physiologic and performance tests used in sports science research and professional practice should be developed following a rigorous validation process, as is done in other scientific fields, such as clinimetrics. If performance tests are used in multicenter studies, they have to be described in detail (e.g., with photographic illustrations), should be demonstrated during onsite visits, and should be controlled during study monitoring.

Exact protocols including detailed test descriptions are required for performance tests. Similar testing procedures should be ensured during onsite visits.

Strength Measurements

Muscle strength tests in typical positions belong to the most common clinical outcome parameters. They not only reflect muscle power but also indicate absence of pain, which enables active force generation. Although they are considered "objective," they underlie a number of influencing factors such as fear of injury, pain, medications, work satisfaction, and other motivational factors with an influence on sincerity of active testing.[225] These factors may vary among study sites depending on cultural background, Workers' Compensation, and other socioeconomic factors.

Another challenge is presented by the variety of measurement devices. For example, shoulder abduction strength, as required for the calculation of the Constant score,[226] can be measured with a number of devices, e.g., spring balance, Isobex (Medical Device Solutions AG, Oberburg, Switzerland), or dynamometer. They all operate on a different working principle and subsequently measure different parameters. If not specified before the study, this may lead to a situation in which data pooling is not feasible. Therefore specification of the measurement device is mandatory in each multicenter study. More information about the measurement protocol is necessary, however, to ensure comparability of data. Positioning of the patient may influence the result. This has been shown for grip strength as well as for hip abduction strength. For example, maximal hip abductor strength is significantly higher in the side-lying position compared with the standing and supine positions.[227] In addition, information about the number of repetitions, as well as further data processing, is required to avoid additional bias. Strength measurements are typically repeated in triplicate. Then, it has to be specified whether the maximum, mean, or median value will be processed according to the research question.[228]

For strength measurements, the exact measurement device, including manufacturer, positioning of the patient, number of repetitions, and selection process of measurements have to be defined to ensure data comparability across study sites.

Sophisticated Functional Tests

More sophisticated functional tests such as in-shoe plantar pressure measurements, gait analysis (instrumented walkway), or force-plate analysis may contribute additional information.[229] For specific research questions, these laboratory methods are considered to be the most accurate measurement methods, and clinicians and scientists tend to include them in clinical trials. For instance, high reliability could be shown for various methods of instrumented foot-function tests.[230] However, a number of issues have to be considered to avoid pitfalls when used in multicenter trials: not only does the technology of the chosen test have to be available at each site, but the know-how to operate it is also crucial. For example, a sophisticated motion-capture system requires skilled staff who can install, calibrate, and run it. Laboratory space, logistics for patient handling, computational resources, and experiences in patient testing are necessary. If such a method is to be used in a multicenter study, exact definitions of the system and of all laboratory parameters applicable to all sites are mandatory, as well as careful training. If the specific laboratory test is not feasible at all sites, it is an option to perform the test in a study subgroup only at clinics with the required infrastructure and resources.

Sophisticated functional tests may provide additional information for a given research question but require specific infrastructure, know-how, and resources. If not available at all sites, these methods can be limited to a subset of selected sites to collect the additional information.

Radiographic Evaluation

Radiographic parameters are part of almost all orthopaedic studies. However, despite the widespread use, only little consensus exists about radiographic grading. Interrater agreement measured with the κ coefficient ranges from 0.4 for sclerosis to 0.95 for joint-space narrowing as shown by Lane et al.[231] This broad range was recently re-emphasized in another study investigating the reliability and agreement of measures used in radiographic evaluation of the adult hip.[232] The authors also stated that direct measurements (femoral head diameter) were more reliable than measurements requiring estimation on the part of the observer (Tönnis angle, neck-shaft angle). Agreement between repeated measurements showed many parameters with low absolute reliability. The same problem was reported from the quantification of fracture classification,[233] reduction,[234] and healing.[235,236]

However, the information of an image is stored in the radiograph. Central radiograph reading may help to extract the required data and to avoid subjective judgment by the treating surgeon on the one hand, and it is more reliable in detecting all suspicious findings and less biased by the surgeon's perspective on the other hand. Establishing a radiology review board for a multicenter study is a worthwhile method to increase data quality.[237] The images should be collected centrally, and a minimum of 2 experienced investigators should evaluate the blinded radiographs independently. It is recommended to collect the digital radiographs in DICOM (Digital Imaging and Communications in Medicine) format for later image processing. Clear definitions of each radiologic parameter documented in the study plan or an image-reading manual is mandatory.[238] An initial training session may help to improve interrater agreement.

Central image reading by 2 independent, experienced observers and consensus finding help to increase data quality. Strict radiologic definitions are mandatory; an initial training session may help to improve agreement.

Bone Density Measurements

Local bone density and systemic osteoporosis status both came into focus in several studies.[239,240] A typical example is the change in local bone density around joint replacements as a reaction to different prosthesis designs.[241] Although many authors refer to predefined areas like Gruen zones,[242] they may vary from group to group depending on the exact definition. In a multicenter study, the measurement method (e.g., peripheral quantitative computed tomography or dual-energy absorptiometry), the exact device, and the imaging parameters, as well as the processing algorithm, have to be specified. Especially the differences between different devices for dual-energy absorptiometry introduce a large source of variability in studies with several study sites. These devices are often calibrated with cohorts provided by the manufacturer only. Therefore pooling of absolute values is unfeasible; only relative spatial or temporal changes can be compared or pooled.[243] Limitation to one device type only reduces the number of potential recruitment sites in many studies.

However, if peripheral quantitative computed tomography is feasible within a multicenter study, cross-calibration with a standardized phantom (e.g., the European forearm phantom) improves data quality[244] and allows pooling of the absolute values. Study protocols including bone density assessment should include documentation of precision accuracy and stability at one site as well as comparisons between different sites. A protocol for the circulation and testing of a calibration phantom helps to ensure the required data quality.

Quantification of local and systemic bone density has to be defined in detail including measurement site and area, measurement device, imaging protocol, and (cross-)calibration.

PATIENT-REPORTED OUTCOMES IN MULTICENTER STUDIES

Patient-reported outcomes (PROs) are subjective parameters that come directly from the patient. In contrast to objective parameters, they should exclusively reflect the patient's health condition (e.g., function of the knee, ability to walk, pain, and HRQL) without any space for interpretation by a clinician. They can be used to obtain information on the actual status of a sign or symptom of the patient (e.g., on the preoperative status of an arthritic joint) or to see changes of a sign or symptom over the time—for example, to assess the effect of a medical treatment or the success of a surgery.

Choosing the Conceptual Framework

The 4 target domains that contribute to functional outcomes can be viewed as physical, mental, emotional, and social in nature.[245] In treating patients with impingement, for example, there is a need to facilitate clinical decisions where surgeons must weigh, either explicitly or implicitly, the expected benefits of a particular intervention, whether surgical, medical, or rehabilitative, against the potential harm and cost.[246,247] The choice of an appropriate disability conceptual framework to classify different domains and instruments is fundamental because there is a lack of consistent language and uniform definitions when defining physical function. However, without a common metric to measure these targets, we would be unable to compare results across trials and guide clinical decision making.

The main purpose of the International Classification of Functioning, Disability and Health (ICF) of the World Health Organization to provide a common language to describe disability concepts has made the framework widely popular.[248] Functioning and disability are described in the ICF in terms of the dynamic interaction between health condition, personal factors, and the environment. The ICF is not only a classification system for the impact of disease, it is also a theoretical framework for a relation between variables. The ICF places the emphasis on function rather than condition or disease. The ICF provides a description of situations with regard to human functioning and its restriction. The information is organized into 2 parts: part 1 deals with functioning and disability, whereas part 2 covers contextual factors. Each part has 2 components: The body component comprises 2 classifications, 1 for functions of body systems and 1 for body structures. Activities may be limited in nature, duration, and quality.[249] Activity limitations are referred to as disabilities and are scaled by difficulties and whether assistance is needed to carry out the activity. The ICF has been identified by the American National Committee on Vital and Health Statistics as the only viable code set for reporting functional status.[250]

The design and conduct of good comparative studies in this context rely on the choice of valid instruments that are reliable and responsive.[251] Should the instrument assessing functional outcomes prove to have good psychometric properties, the value of the published literature would be enhanced.[252] However, pragmatic qualities such as the applicability of such instruments in trials examining specific populations, for instance, femoroacetabular impingement and hip labral pathology, should also be considered in addition to the psychometric properties. For example, logistical choices for use of functional outcome instruments should take into consideration the burden to administer, require additional training, and have an adequate score distribution as well as format compatibility.[253] To obtain comparable results, it is necessary that all participating centers use the same version of an outcome measure and perform it in the same way (e.g., direct distribution or telephone interview). This is especially important for those instruments where different versions exist, e.g., the HRQL instrument SF-36 (version 1 or 2, 1 week's recall or 4 weeks' recall) or the Constant score at the shoulder.

- *Use a framework to classify health concepts, whether impairment or activity participation.*

- *Use both disease-specific and generic health measures.*
- *Use instruments with tested psychometric properties.*

Cross-Cultural Challenges

The cultural background can be an important confounding factor in international multicenter studies and also in national studies including migration populations of different cultures. For example, illness behavior and perceptions of pain are different between Americans and Asians.[254]

Have you ever thought about how to assess the same item in patients from different countries, with different cultural backgrounds and different functional demands?

For example, East Asian people use different functions of the hand when eating with chopsticks than Western people. In many cultures, kneeling is an important function regularly practiced during eating or praying, with highest functional demands because of the maximum flexion of the knee.

When using a PRO, it is important that it is available in the national language of the target population because it should be answered by the patient in context with his or her cultural background. Availability in another language does not mean that it has simply been translated by one interpreter or even by a doctor during the interview with the patient. An instrument that should allow reliable comparisons with other studies (e.g., comparing treatment effects) or will be used in an international multicenter study should undergo a careful methodologic process of cross-cultural adaptation and validation such as or comparable to the process described by Guillemin et al.[255] and Beaton et al.[256] (Fig 12). The questionnaire must be correctly translated not only for all questions and answers but also for all instructions for the patient and for the scoring method. For all steps of such a process, careful written documentation that highlights difficulties to reach equivalence between the original questionnaire and new-language questionnaire is necessary.

The first step is the translation into the target language. This should be done independently by 2 bilingual translators with the target language as their mother tongue. One of the translators should be aware of the concept of the questionnaire and should have a medical background. The second translator should have no medical background and be uninformed regarding the concept of the questionnaire. Both translators produce 2 forward-translations, versions T1 and T2.

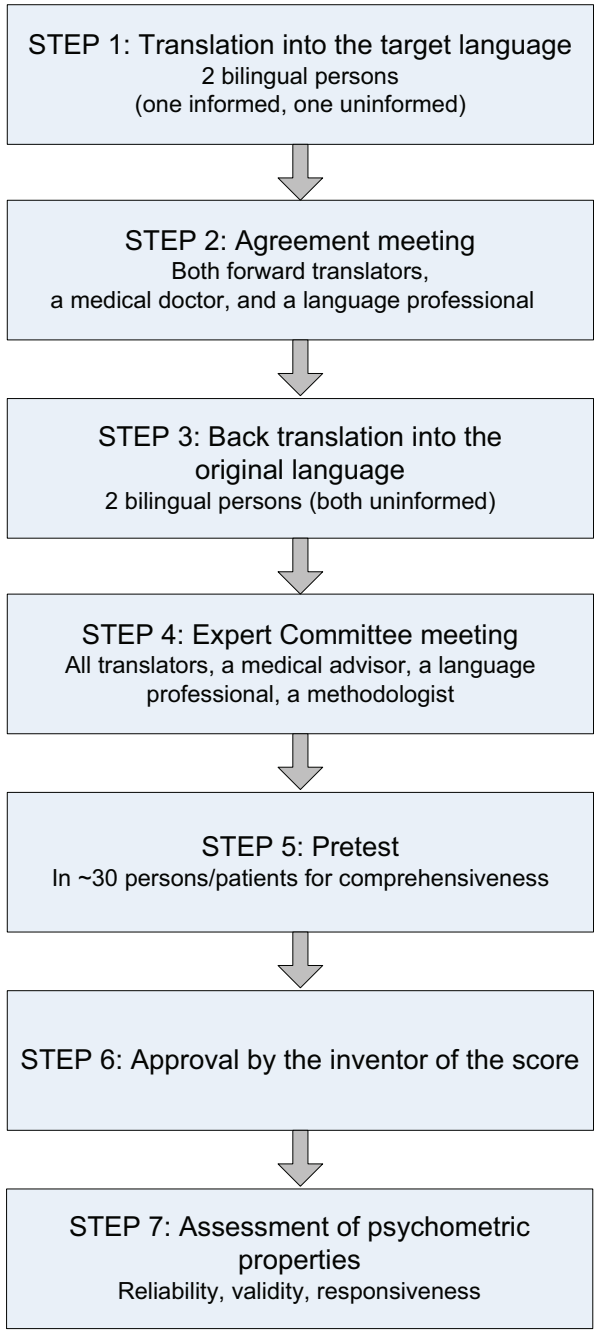

FIGURE 12. Steps of cross-cultural adaptation. Adapted from Beaton et al.[256]

The second step is an agreement meeting, where both forward-translators find an agreement on the translations and produce a synthesis version (T12). The discussion should be led by a third person acting as mediator, e.g., a medical doctor familiar with the questionnaire and its concept. A language professional

such as a linguist is highly recommended for this meeting to help in the discussions of whether equivalence between the original version and new-language version has been reached. All decisions should carefully be documented with their reasons.

The third step is the back-translation of the synthesis version T12. This back-translation should be performed by 2 bilingual persons with the language of the original questionnaire as their mother tongue. They should have no medical background and no knowledge of the original questionnaire and its concept. Both back-translations should be performed completely independently; they produce versions BT1 and BT2. The back-translations can reveal ambiguous wording in the translation or that the translation was incorrect.

The fourth step is the expert committee meeting that finally discusses the synthesis version, as well as the back-translations, and its equivalence to the original questionnaire. All forward- and back-translations and the original and synthesis versions should be available at this meeting. The expert committee should preferably consist of all 4 translators, 1 language professional, 1 medical advisor, and 1 methodologist. They discuss whether the synthesis version can be used for the next steps or whether some phrases need to be changed because of obvious findings from the back-translations. A pre–final version will be created.

In step 5, this pre–final version is tested in 30 subjects/patients for comprehensiveness. This test should not only include the translated questionnaire but also a section where the subject can describe difficulties in understanding or give an idea of his or her interpretation. The results of this test, all documentary material of the process, and all versions should be sent to the inventor of the score for approval of the new-language version.

After approval, the psychometric properties of the new-language version must be assessed, i.e., it must be tested for reliability, validity, and responsiveness.

1. The test-retest reliability is important to determine whether the score is stable, i.e., the score does not change if the patient's health condition has not changed. It is, for example, tested in a symptom-stable interval where the patient has to complete the new questionnaire twice within a few days (usually not more than 1 week).
2. The validity must be tested to determine whether the new score measures what it is supposed to measure. For example, a new score of the hip can be tested together with an HRQL instrument such as the SF-36 or with a region-specific instrument such as the WOMAC. If 2 instruments correlate well with each other, this indicates convergent validity, and if not, there is divergent validity between the scoring systems. For a new score of the hip, we would expect high (convergent) correlations with the WOMAC and the physical function subscales of the SF-36 but low (divergent) correlations with the mental subscales of the SF-36.
3. Testing responsiveness is necessary to determine whether the score detects changes of a symptom or health condition over time or after a treatment.

Only if the new-language score performs well on all these tests is it deemed a sufficient and reliable instrument.

A different cultural background can be an important confounding factor. If a patient self-assessed questionnaire shall be used in another language, it cannot be simply translated. It must undergo a careful process of cross-cultural adaptation and testing.

Influence of Comorbidity

Because different and severe comorbidities exist in older patient populations, clinical results may not be representative of all patient types presenting with knee problems. If patient-oriented measures are used only for healthy and lucid patients, PROs will be ineffective in the 31% to 45% of patients in geriatric orthopaedic rehabilitation units who are reported to have cognitive impairment.[257] Patients who fall into this category are often challenging to deal with because of a lack of compliance or mortality rates that lead to a loss in follow-up. RCTs should diagnose comorbidities and screen patients for either inclusion or exclusion according to the research question and monitor higher degrees of disability through functional outcome assessment. Impaired physical function has been linked to many indicators of increased health services utilization and has become fundamental for researchers, clinicians, and funding agencies.

We can gain greater understanding of the patient's perspective by using the appropriate instruments that measure all aspects of functional recovery. Health in elderly patients is often compromised by various comorbidities of differing levels of severity. These areas need to be further investigated with the aim of finding a common metric to assess specific populations implicating a change in the future conduct of EBM. The implications of obtaining results that are not representative of

all patients are major, and future research needs to investigate the potential use of proxy responders.

The Use of Proxy Responders: Ethical Considerations

Cognitive impairment does not necessarily lead to an assessment of incompetence. Cognitively impaired elderly persons may still be able to make morally responsible decisions, based on different degrees of decision making and on their personal history.[258] The solution is more complex, and assessment instruments need to be used with caution when dealing with multiple interactions between medical personnel and patients and their family members. With regard to the ethical considerations and care of patients with dementia, the approach needs to encompass the patient's feeling of being cared for and approached as a competent or free individual. Because the elderly are regarded as a vulnerable group, it is important to use a method that can protect those who cannot decide, as well as to provide the opportunity to participate in research for those who are able to decide for themselves. Especially because the law prohibits scientific research on incompetent patients, unless special conditions are fulfilled, a close investigation on the issue of informed consent is needed and should be recommended for future research. Patients with mild to moderate dementia still have moral capacity.

Proxy and patient responses are not interchangeable; however, proxy responses can provide an option for assessing function and health status in patients who are unable to respond on their own behalf. In a prospective longitudinal study examining agreement between patient and proxy respondents using the Health Utilities Mark 2 and Mark 3 (Health Utilities Inc., Dundas, Ontario, Canada) over time during a 6-month recovery after hip fracture, the authors reported ICC values from 0.50 to 0.85 ($P < .001$) for physically based observable dimensions of health status and from 0.32 to 0.66 ($P < .01$) for less observable dimensions.[259] Future investigation of the proxy interrater agreement with the use of health status instruments is needed.

GENERAL CONSIDERATIONS FOR MULTICENTER STUDIES

The homogeneity of observations throughout the whole study group plays an important role, as does the absolute number that can be achieved within a given time frame. Whereas clear inclusion and exclusion criteria combined with well-defined outcome measures help to limit variability, these factors also limit patient recruitment and generalizability.[260] Therefore extreme unification of variables and criteria may lead to problems in statistical power. Nevertheless, a number of boundary conditions should be defined to limit variability. They include the following:

- preoperative patient preparation, e.g., positioning
- standardization of surgical intervention including approach, concomitant interventions, e.g., soft-tissue release or any tenodesis, and wound closure
- perioperative antibiotic protocol and subsequent infection prophylaxis
- anesthesia and postoperative pain management
- thrombosis prophylaxis
- postoperative rehabilitation protocol including timing and extent of passive and active range of motion, casts, and weight bearing.

Illustrations, flowcharts, pocket charts, and checklists are helpful to achieve a similar information level at all sites. It is of paramount interest to integrate the surgical staff as well as colleagues from anesthesiology in the information flow. Only if all persons involved in the treatment process are informed and act according to the study protocol can the quality of outcome measures be ensured. For example, an incorrect anesthesiology protocol may lead to patient exclusion if it possibly interferes with an outcome measure such as postoperative pain. All changes to the protocol have to be communicated as protocol amendments to all persons and the appropriate institutional review boards.

Newsletters to all sites and all collaborators are good instruments to ensure a similar information level among all sites and collaborators at one site. They may also motivate partners for active patient recruitment and subsequent adherence to the study protocol (Tables 37-38).

TABLE 37. *Multicenter Checklist*

- Reach an agreement within the study group about the exact protocol of intervention and related treatment (preoperative patient preparation, anesthesia, pain management, thrombosis prophylaxis, postoperative rehabilitation)
- Ensure compliance with the protocol at all sites
- Define measurement devices and units
- Define time points for follow-up including tolerance
- Consider cross-calibration of devices, e.g., with phantoms
- Define each (!) variable including measurement procedure

TABLE 38. *Tips and Tricks*

- Teach the study nurses of all sites at the beginning of the study
- Train the investigators in the correct performance of objective tests
- Provide pocket flowcharts for patient monitoring
- Use newsletters to keep all sites updated
- Publish the study protocol, e.g., at www.clinicaltrials.gov

TABLE 39. *Feasibility Questionnaire*

- How many cases of the study disease/injury (according to inclusion criteria) do you have per year?
- How many cases (%) of the study disease/injury (according to inclusion criteria) come to your clinic for follow-up examinations on a regular basis?
- When do your patients usually come to follow-up examinations with the study disease/injury?
- How many cases (%) of the study disease/injury (according to inclusion criteria) do you expect to come for the follow-up examinations planned for the study?
- In case patients could not come to a follow-up, could you perform telephone interviews?
- Do you treat your patients according to the study protocol?
- Do you perform the same postoperative treatment/rehabilitation?
- Which devices/techniques do you use (specific question, e.g., for densitometry, strength measurement, gait analysis)?
- Do you have a study nurse or dedicated person who could run the study and manage the operational affairs?
- How long do your institutional review board submissions usually take until approval?
- Do you prefer electronic data capture or paper-based CRFs?
- Can you perform and send radiographs in DICOM (Digital Imaging and Communications in Medicine) format?

Data Acquisition

For participation in a multicenter study, it is not enough to treat patients who potentially might be included in the specific study. A few more conditions are mandatory to act as a study site and to successfully contribute observations. If surgeons agree to participate in a study, they should be aware of the necessary infrastructure. This includes a person who will collect the data. In the best case, this is a well-trained study nurse who is responsible for all study performance–related issues at one or more clinics. Experiences with high-quality study nurses show that they ensure not only completeness of data but also a high follow-up rate. A local scientific contact person with a key interest in the study subject also helps to get a study locally established. This person can be instrumental in the necessary application to the institutional review board or local ethics committee. In addition, he or she has to inform all necessary partners at the study site, such as anesthesiologists, members of the operating room staff, and physiotherapists.

A central infrastructure for data collection is required because of differences in local infrastructure. Whereas in most single-center studies the clinical information system can be used, in multicenter studies additional data collection is mandatory. The system used has to be tailored to the specific needs of the study (e.g., acquisition of image data in contrast to patient self-assessment) but also to legal boundary conditions such as data safety, query management, or guidelines of good clinical practice.[261] In many studies paper-based data acquisition with subsequent telefax transmission is still a valid option. Web-based databases are an interesting alternative. They have the advantage that data validation and query management can be implemented electronically, thus improving data quality, but more time is usually required to insert data and to comply with data safety issues. So far, no general applicable gold standard for data acquisition exists, but this issue has to be addressed during the planning phase to avoid missing data or data loss. It requires a separate budget for programming and maintaining the database and for data processing, if applicable.

CONCLUSIONS

Multicenter clinical studies offer a unique chance to obtain high numbers of patients even for rare diseases or fractures. They are of higher value than single-center studies because they show a more real-world approach and not only the high quality of a well-skilled orthopaedic surgeon in an ideal clinical environment. The great opportunities of a multicenter clinical study sometimes induce the planning clinicians to assess as much as possible, i.e., numerous PROs and objective measures. However, the best clinical study is only as good as it is feasible. Therefore a feasibility questionnaire in the planning phase of a multicenter study is often helpful (Table 39).

Finally, it is of utmost importance to carefully plan finances for managing multicenter studies!

Sabine Goldhahn, M.D.
Amy Hoang-Kim, M.Sc., Ph.D.(cand)
Norimasa Nakamura, M.D., Ph.D.
Jörg Goldhahn, M.D., M.A.S.

SECTION 17

Reporting of Complications in Clinical Trials

Complications in orthopaedic trials are an essential source of information. They may result in discontinuing unsuccessful treatment strategies, help to identify potential for development, and form the basis for shared decision making with patients. However, the term "complications" implies different things to different people. For surgeons, they seem to cause trouble in the first instance. In addition, they impair any success rate, they may need re-intervention, they often require extensive communication with patients, they sometimes lead to legal problems, and they are, all in all, associated with more problems and high costs. Given their perception as failures, it is not surprising that some surgeons tend to neglect them—at least in terms of reporting them. Other surgeons are more critical, and they document and report more complications. So, what is regarded as a complication is dependent on the surgeon's understanding and awareness. The great variability of reported complications for specific indications illustrates this fact. A recent survey among orthopaedic surgeons supports this observation by showing different awareness levels of complications.[262] Herein, we suggest a standardized approach to documenting, assessing, and reporting complications in clinical trials.

COMPLICATIONS FROM DIFFERENT PERSPECTIVES

For legal authorities, complications are so-called adverse events that must be reported according to the guidelines of good clinical practice. They are interested in information—for instance, whether the complication leads to death or another stay in the hospital (serious adverse event) or whether it is device related.[263] Reported complications may lead to a study being stopped or implant withdrawal from the market or may have legal consequences.

For patients, complications mean a decrease in quality of life in the first instance. A treatment may take longer than usual, may cause more pain than expected, may result in a poorer result than promised, and may lead to long-term sequelae. It could also result in a re-intervention to correct these conditions or to prevent long-term consequences. Primarily, patients are not interested in the surgeon's perspective or in the legal perspective. They simply want to have function and quality of life re-established, and they regard everything that deviates from the normal course of healing and rehabilitation as a complication. In addition, they should obtain unbiased information about expected complication risks as a basis for shared decision making.[264]

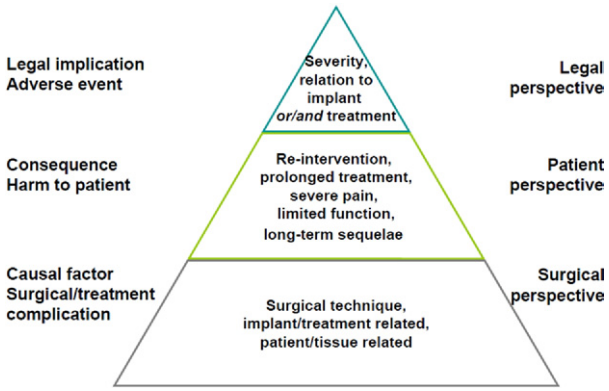

FIGURE 13. Hierarchy of complications. The pyramid illustrates the hierarchy from causal factor to patient harm until legal adverse event classification. This corresponds to the different perspectives on the right side of the diagram.

The outlined consequences show that it seems almost impossible to satisfy all perspectives at the same time. Therefore a pragmatic approach is required that should acknowledge relevance. A severe complication may lead to a decrease in surgical reputation and/or withdrawal of an implant with some financial consequences for the manufacturer. However, a patient may suffer from consequences of complications for the rest of his or her life or even die. So, complications have the highest relevance for the person experiencing them. Therefore definitions of complications have to be patient centered.

This leads to a hierarchy of complications (Fig 13). Whereas the surgical perspective is based on experience and always includes reasoning and causality, the patient perspective serves as a filter. Any event without any harm or consequences to the patient might not be considered as a complication.

On top of the hierarchy, the legal perspective determines the relation to any tested implant or treatment and classifies the severity according to

established guidelines. It is mostly a subset of the complications that may matter to the patients as described above.

In accordance with the good clinical practice guidelines [E2A and E6(R1)] of the International Conference on Harmonization, a serious adverse event is clearly defined as any untoward medical occurrence that

- results in death
- is life-threatening (it should be noted that the term "life-threatening" in the definition of "serious" refers to an event in which the patient was at risk of death at the time of the event; it does not refer to an event that hypothetically might have caused death if it were more severe)
- requires inpatient hospitalization or prolongation of existing hospitalization
- results in persistent or significant disability/incapacity
- necessitates medical or surgical intervention to prevent permanent impairment to a body structure or a body function
- leads to fetal distress, fetal death, or congenital abnormality or birth defect.

CASE EXAMPLE

The importance of the patient perspective as a filter should be demonstrated with the following example.

In the treatment of an unstable trochanteric fracture using a dynamic hip screw, the screw was misplaced very close to the articular surface. The patient claims to have severe pain during weight bearing.

- Surgical perspective: the complication is a screw cutout. Possible causes can be initial misplacement (surgical technique) and/or poor bone quality (patient/tissue related).
- Patient perspective: the patient has severe pain and reduced function and may have long-term consequences if untreated or will face a re-intervention to prevent them.
- Legal perspective: the severity classification depends on the possible re-intervention. The possible relation to the implant depends on the judgment of the surgeon in terms of whether the malpositioning was related to poor surgical technique and/or device.

The case example demonstrates different issues: (1) The patient suffers under all circumstances regardless of the causative factor or the legal classification. (2) The surgeon can influence the classification of adverse events, e.g., by accepting poor functional outcome or neglecting re-intervention.

NORMAL EXPECTED COURSE OF HEALING

If the patient perception of a complication is any deviation from the normal course of healing and rehabilitation, then a definition of "normal" is required. Healing of any tissue such as bone, cartilage, or tendon has a broad range depending on patient characteristics as well as on the specific intervention. For instance, time to fracture union is not clearly defined and depends on many confounding variables and on the assessment method.[265-267] Therefore thresholds are required that distinguish the normal course from a pathological course of healing. The same is valid for pain and return to function.

Whereas a certain amount of pain caused by wound and tissue healing after a surgical intervention is related to the normal course of healing, prolonged pain has another cause in most cases. The same is valid for return to function and activities of daily living. A certain improvement of function with a wide range is expected at given time points after intervention. However, complete loss of function or significantly lower function than expected and subsequently impaired activities of daily living have to be considered complications.

Thus, for both pain and return to function, thresholds have to be determined for the normal expected course of healing. Everything outside of these has to be considered as a complication or the consequence of a complication. Pain and low function are often only symptoms of an underlying, often anatomic problem (e.g., articular step or valgus deformity). If patients report severe pain and/or limitation of function, it is necessary to search for the causative problem.

COMPLICATION REPORTING

For each study, the normal course of healing and rehabilitation including an evidence-based range should be defined. This includes pain and functional status at each follow-up, as well as healing of any investigated tissue such as cartilage or bone.

Anticipated complications/adverse events should be listed in all study protocols with clear and objective definitions along with appropriate scientific references.

TABLE 40. *For Each Complication, a Minimum Set of Information Should Be Documented Because of Regulations and to Allow Clinically Meaningful Evaluation and Reporting*

Domain	Variables
Identification	*1.* Investigator's name and phone number
	2. Study name
	3. Patient identification (trial number, initials, age, gender)
Treatment	*4.* The treatment number (if applicable, such as in a randomized clinical trial)
	5. The name of the suspect medical product and date of treatment
	6. Product serial number (in case of SADE)
Complication	*7.* Complication type
	8. Date of occurrence or onset
	9. Short description (open text field)
Action(s)	*10.* Subsequent action taken (e.g., operative)
Outcome(s)	*11.* Outcome of the complication at the time of reporting (or end of the study)
Assessment	*12.* Seriousness of the event
	13. Most likely causative factor, e.g., relation to the surgical intervention or the implant used; we recommend using the 4 categories presented in this chapter.

NOTE. This is the minimum information to be collected by means of an adverse event form/complication CRF to be adapted for each study. Investigators are asked to fill in 1 form for each complication; however, more than 1 event may be recorded on the same form if they occurred simultaneously and were unambiguously causally related.
Abbreviation: SADE, severe adverse device effect.

It is important to quantify the standard complication rate known from the clinical literature, the common salvage procedures, and the final outcome that can be expected.

- *For each complication, a minimum set of information should be documented because of regulations and to allow clinically meaningful evaluation and reporting.*
- *In clinical research these variables should be presented as a standard adverse event/complication case report form (CRF) that is adapted for each study.*

In Table 40 minimal requirements for record keeping of complications in clinical studies are listed. Investigators are asked to fill in one form for each complication; however, more than one event may be recorded on the same form if they occurred simultaneously and were unambiguously causally related (e.g., an implant failure simultaneously with a loss of reduction).

Because complications occur as part of a more or less complex chain of events, a clear distinction should be made between the complications/adverse events themselves and the following, as illustrated in Fig 14: their causal trigger factors, their treatment (which could actually be no action), and their consequences or outcomes.

FOLLOW-UP OF COMPLICATIONS

- *If an original complication record states that the complication was resolved or that the recovery process is completed (with or without damages), no further data are required.*

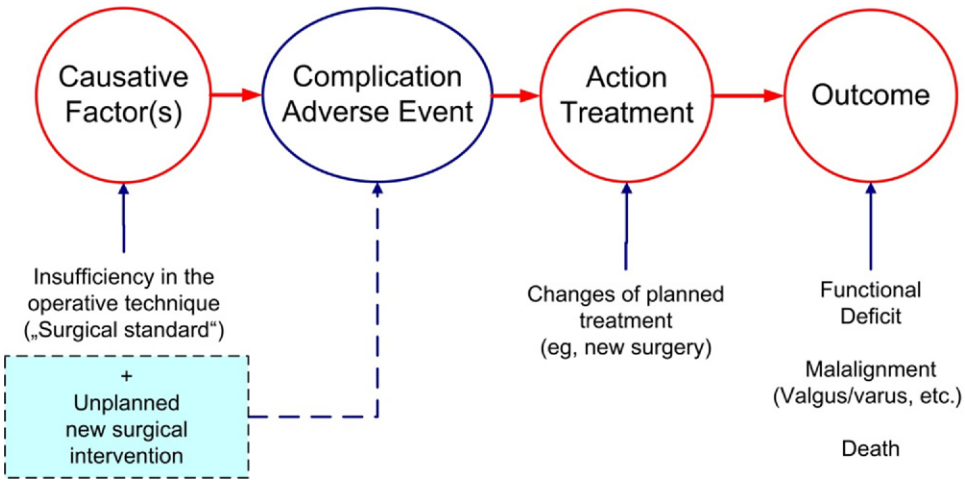

FIGURE 14. Clear distinctions should be made between complications/adverse events themselves, their most likely causal factors, their treatment (which could actually be no action) and their consequences or outcomes.

TABLE 41. *Classification of Complications*

Category	Class	No.	Example
Treatment related	Related to surgical technique	1a	Malpositioning of screws, wrong procedure
	Related to device/treatment	1b	Loosening of polyethylene glenoids due to wear
Patient related	Related to local tissue condition	2a	Cutout of correctly placed screw due to poor bone quality
	Related to overall patient condition (e.g., systemic)	2b	Myocardial infarction

- *Alternatively, it is necessary to follow up the complication until it is resolved, in particular in terms of its treatment, outcome, and assessment, and all new information must be documented.*

In clinical studies a follow-up adverse event/complication CRF should be distributed to investigators to capture this information until complications are resolved or finally evaluated at the end of the study.

CLASSIFICATION OF COMPLICATIONS

We propose 2 main categories, treatment related and patient related, and 2 subsequent classes for the classification of complications based on their most likely causative factor (Table 41).

Of course, many cases remain where the causal relation is a topic of debate. For instance, it is still not clear whether an avascular head necrosis is the result of the surgical treatment of a humeral head fracture or would correspond to the normal course of disease.

However, careful planning combined with prospective definition of complications and their causal relation increases the study quality. This planning phase may lead to an extensive list of anticipated complications, as shown in a recent trial,[238] but helps to categorize complications before the study and will result in an unbiased complication analysis at the end of the study.

DATA QUALITY CONTROL AND FINAL REVIEW OF COMPLICATIONS

Active monitoring and quality control are essential to avoid or limit under-reporting and misleading complication results. To favor completeness and correctness of documentation of complications, 1 or more of the following measures can be implemented in any study:

1. Source data verification during monitoring visits
2. Active reporting: implement systematic assessment of any complication at each examination visit (e.g., using standard CRF or asking whether another physician was visited other than for routine assessment)
3. Incentive to report: facilitate simple recording process and ensure anonymous reporting of complication statistics outside the involved clinics so that results cannot be traced back to the individual treating surgeon
4. Obtainment of additional information on putative events from the patient's family doctor, if necessary
5. Evaluation of reported complications by the study's principal investigator, an independent experienced clinician, or any specifically established complication review board (CRB)

The final complication review should be conducted based on complication/adverse event forms, as well as available related diagnostic images, which might be made anonymous to limit bias. Complication data are reviewed for their clinical pertinence, classification, and severity, as well as relation to the investigated treatment or medical device. All changes and data corrections should be thoroughly justified and documented.

ANALYSIS OF COMPLICATIONS

A minimum set of complication analyses should be conducted in any study. However, it should be noted that if regulatory requirements oblige investigators to document all complications occurring during a study, only a specific clinically relevant subset may be analyzed to answer a study objective. It is critical to clearly define which complications are included in such a subset and to specify the period of observation (e.g., intraoperative, postoperative, and follow-up periods) to allow appropriate interpretation of the results. In the context of prospective clinical investigations, the timing of observation for each patient starts with the initiation of treatment or primary surgery and

TABLE 42. *Example of Presentation of Complication Risks*

Type of Complication	n*	Risk (%)†	95% Binomial Exact Confidence Interval
Postoperative local implant/bone complications	18	10.2	6.1-15.6
Implant			
Blade migration	1	0.6	0.01-3.1
Implant breakage	3	1.7	0.35-4.9
Cutout	2	1.1	0.14-4.0
Other implant complications	2	1.1	0.14-4.0
Bone/fracture			
Loss of reduction	1	0.6	0.01-3.1
Neck shortening	8	4.5	2.0-8.7
Other bone complications	6	3.4	1.3-7.2

NOTE. The number of patients (N) equals 177.
*Number of patients with at least 1 complication (meaning that the patient can have >1 complication, but for the risk calculation, the number of patients having complication[s] is used).
†Number of patients having a specific complication divided by the number of patients being enrolled in the study.

ends at the end of the study. For the report of complications, complication risks can be calculated and presented as shown in Table 42.

Complication risks should be presented based on the number of patients experiencing complications and not the total number of documented complications.

POINTS TO CONSIDER

According to our experience, for many surgeons, an event that is unrelated to the treatment may not be considered as a complication and must therefore not be documented. In addition, clinicians may sometimes believe that they do not need to document events that have no or limited consequences for the patients to avoid documentation overload. Nevertheless, harmonized standards for the conduct of clinical trials define a complication as "any untoward medical occurrence" not necessarily related to potential causal factors or severity[268]; even mild anticipated complications in the framework of clinical research require official reporting to authorities if their rate of occurrence is higher than what can reasonably be expected in any study. Whereas all complications must be documented from a regulatory viewpoint, the primary analysis can be focused on the patient-relevant complications.

At the end of a clinical study, complications/adverse events should be assessed and discussed by a complication review board in a complication assessment meeting. The completed complication case report forms, additional documentary material, and all images of the patients should be available for such a meeting.

We want to stress the importance of conducting an independent review of complications for the credibility of safety data. Complication rates in the literature are most often elusive.[262] In addition, they are likely underestimated, in particular when documented by the inventor(s) of any surgical technique. Despite all efforts at standardization, the assessment and reporting of complications will always require clinical judgment and therefore remain partly subjective. A CRB can address such limitations and, in our opinion, can also be established for single-center studies at a low cost. A CRB can, for instance, consist of 2 to 4 orthopaedic surgeons (at least 1 of whom should not be involved in the study), a radiologist, and a methodologist. It is to be distinguished from any data monitoring committee[269] established as part of large multicenter studies; whereas the CRB is set to control the relevance and integrity of the complication records, the data monitoring committee is set to review the occurrence of complications (i.e., assess the validated data and decide on the continuation of a study). The primary role of the CRB as we propose is to perform quality control and consolidate complication data before their analyses.

Sabine Goldhahn, M.D.
Laurent Audigé, D.V.M., Ph.D.
Norimasa Nakamura, M.D., Ph.D
Jörg Goldhahn, M.D., M.A.S.

SECTION 18

How to Write a Scientific Article: A Painless Guide to Putting it All Together

Before we get started, we have to say it: it takes a great deal of discipline to complete a scientific article. So many authors start; so few finish. But, most important of all is to start. If you don't start to write, you will never finish. If you just start and get going, it is often easier than you thought in the beginning.

DISCIPLINE

Have you ever heard of the 80:20 rule? Some argue it takes 20% effort to get 80% of the results. This is not true when it comes to writing a scientific article. We introduce a different rule; we call it the 90:10 rule. With average effort, an author can complete 90% of the paper. However, it takes just as much time to do the last 10%. Be prepared, because the last 10% often is the part that makes a difference. So we tell you in advance that finishing a scientific article takes time and discipline.

PASSION

Young ambitious authors sometimes do research for no other point than just doing research. Some do research to boost their career or because their work requires it. Perhaps the research is a requirement of an educational program, or perhaps the author is ego driven and wants to make a contribution, add to his or her curriculum vitae, or just get involved in something new. The problem is that such authors are at risk of choosing a research topic about which they lack passion. Then only pure discipline becomes the motivation for driving to completion of a dispassionate paper. Such authors are unlikely to continue a long research career.

Experienced and successful authors write about topics for which they have passion. You might say that brains are good, but passion is better. They get into a flow and putting it all together is a pleasure. The trick is to select a research topic for which the author has passion and interest. This shouldn't be difficult, because the purpose of a research study is to answer a question. Therefore, if authors really have a question, then they really want to know the answer. In other words, they like to make a difference.

If an author really wants to know the answer to a question, then he or she is by definition interested in the topic. Passionate interests in finding answers to clinically relevant questions prevent authors from becoming bored with their research before they get anywhere close to completing their projects.

SELECTING THE CLINICALLY RELEVANT QUESTION

Start by selecting the journal in which you are interested in publishing. Read that journal. Collect that journal. Then, when you're ready to start a research project, sit down and go through the last year or 2 of all the relevant articles. But don't overdo it; you don't have to read everything. Take your time and select the relevant papers; not more. Already at this stage and over and over again throughout the entire project, stay updated. Don't rely on old references only. The classical ones are still good, but things happen fast.

Read the articles in which you are interested in. Read the discussions of those articles, and at the end of the discussion, before the conclusion, search for the limitations of the study. The authors should have spelled them out. In fact, an honest report of limitations is often what brings the science forward. It creates new interest and poses new questions. So when you start to write yourself, don't forget to state the limitations of your own study.

In the limitations portion of the discussion, good authors suggest future research that will be necessary to address current limitations in the medical literature. Good readers should be able to think of other limitations of the study and future research to address those limitations. Taken in sum, this is how to choose a topic.

To review: read the literature, find the limitations, list future research to address the limitations, *and perform this future research as your new project.*

PURPOSE

The purpose of your study is to answer a question. To review the section on passion above, make sure it's a question in which you're really interested in finding the answer. Otherwise, you may risk never finding the answer.

HYPOTHESIS

Don't wait. The best research is prospective. Before you start your study, choose your hypothesis; it doesn't matter whether the hypothesis is right or wrong, because your research is to test this and you will find out the answer later. The hypothesis is what you think the answer to your question will be before you start the study. In other words, what do you expect to find or prove with your study?

LEVELS OF EVIDENCE

First, familiarize yourself with the tables summarizing levels of evidence.

Remember, editors prefer original scientific articles of the highest possible evidence level. Sometimes, they get what they want, but too often they don't.

Level V evidence (expert opinion) is the lowest level.

Level IV is a case series and is also low level of evidence. Unfortunately, while a case series can be of value, case series are the most common in the surgical literature. The problem is that case series do not include a control group.

Level III evidence is retrospective comparative research. Comparison of a control group is excellent, but prospective study is better than retrospective study. Strict inclusion and exclusion criteria mitigate against selection bias. There are several types of biases, but selection bias is probably most common and can easily skew results.

Level II evidence is prospective comparative research. However, this method is not randomized, which can result in selection bias. Strict inclusion and exclusion criteria mitigate against selection bias.

Level I evidence is a prospective randomized controlled trial. Randomization mitigates against selection bias. Level I evidence is the highest level of evidence.

Studies of higher levels of evidence are required to compare the effectiveness of one technique versus another technique. Chances of acceptance of an original scientific article are increased for studies of higher levels of evidence.

INTRODUCTION

The body of the introduction frames the question you will be asking. The purpose and hypothesis of your study should be stated at the end of the introduction.

Warning: no one is going to read your paper if the introduction is boring. Therefore, the introduction should not be overly long. Further warning: you're not going to want to finish writing the paper if your topic is boring. Therefore, stop right here, and go back to the purpose section above, and choose a more controversial and interesting question. Controversy in science is good; don't be afraid of it.

The good news is that you have selected a topic that is not boring. Good job, and congratulations because your introduction will draw the reader into reading your paper. One good thing leads to another and at the end of the day, your paper will be cited by other researchers, because it was passionate and not boring. People will bother to read it.

Don't waste the reader's time. Make your introduction short and highlight the controversy.

METHODS

The methods should be reproducible. Other researchers must be able to copy what you did. We teach this the same way every time: the methods should be like a cookbook. Give a step-by-step description so that your study can be repeated by other authors. It should include clear inclusion and exclusion criteria. Everything that you plan *before the study starts* should be a part of the methods.

At the end of the methods, include a description of the statistics that you will use to analyze your results. If it is a comparative study, be sure to include a power analysis to determine the cohort size. This is far too often missing or incorrectly done. A power problem is probably the most common problem in the majority of clinical studies. One might say that an overwhelming number of clinical studies are underpowered and therefore not conclusive. A study should include enough patients, not fewer than needed and also not too many. It is unethical to perform a clinical study on a new (possibly experimental) surgical technique, and either underpower or overpower the study.

Therefore, before you start your study, hire a statistician. *Hire a statistician before you start your study*. This is so vital that we said it twice!

People sometimes ask us, where do you find a statistician? They tend to work at universities, especially in those with research departments. If you don't work at a university, we would suggest that you contact the nearest department of public health to solve the problem. In addition, sometimes industry partners may employ statisticians whose non-commercial interests include research and education.

Since we're not statisticians, we really tried to keep this simple. Let's just focus on the most common statistical mistake.

The number one statistical mistake is doing a study that has too few patients. This is a study with inadequate power.

Underpowered studies mean authors might show no difference between two groups. And the results could be wrong. If there is not an adequate number of patients in each group, the authors could be making an error (we call it beta error) and this is probably the most common error in clinical studies.

The good news is that you performed a power analysis and it determined the minimum number of patients you need to avoid beta error. What do you do next to determine how many patients are required in each group? Once you know the minimum number of patients, add 25% more. Why? To mitigate against transfer bias, which is loss of patients to follow-up.

Maximum acceptable transfer bias is 20% of patients lost to follow-up. This is an arbitrary journal standard of the highest threshold of transfer bias accepted at 2 years follow-up. Some journals consider that a "worst case scenario," i.e., any patient lost to follow-up is disappointing.

Before you start your research, make sure you have a mechanism in place to ensure staff and research support funding so you can build a team with the ability to achieve a high number of patients completing follow-up after 2 years. Multiple ways to find the patients (friends and relatives) placed in the database will help. A researcher does have to work very, very hard to follow patients over the long-term. You must be patient because a disease has a tendency to disappear when a randomized study is started.

If you do find a difference between 2 groups, then you can determine if that difference is or is not statistically significant.

We test statistical significance with *P* value reporting. By convention, $P < .05$ means that there is only a 5% chance of a statistical finding of significant differences between groups occurring by chance. Confused? If $P < .05$, then a finding of a difference between two groups is probably a correct study conclusion. However, statistical significance does not equal clinical significance and this is something you must always bear in mind.

CLINICAL SIGNIFICANCE

Statistically significant differences between groups must be distinguished from clinically relevant differences. The *P* value does not measure clinical significance. Researchers are usually happy if their *P* value is significant, but their patients may not be equally happy. This possible discrepancy must always be taken into account.

What really matters is overlapping confidence intervals. Overlapping confidence intervals suggest no clinically significant difference between groups, and this represents clinical relevance.

We're not statisticians, and researchers must find a statistician to help them calculate confidence intervals. Or, maybe there is, or will be, some new computer app that will allow new research for us to calculate confidence intervals. Either way, authors must learn to take a careful look to see if there is numerical overlap between the confidence intervals to determine if there are clinically significant differences between 2 groups being compared.

If the confidence intervals overlap, results may not be clinically significant (even if they are statistically significant).

RESULTS

Results include everything you have found after you started the study. The most efficient way to display the data is usually to put your results in tables.

In the text, focus on highlighting the most important results. Present the details in the tables and do not repeat each and every detail in the text. Repetitions are never helpful and never make a manuscript better, only longer.

Everything mentioned in the methods should be noted in the results. Everything noted in the results should be mentioned in the methods.

DISCUSSION

A great first sentence for your discussion is "Our results demonstrate . . ." or "The most important findings of our study are . . . ." Obviously then, briefly

summarize your most important results. Then compare your results with the published literature. Then, contrast your results to the published literature. If your results do contrast, try to explain why. This is most probably the highlight of your paper. After you compare and contrast your results to previously published literature, remember to state study limitations before the study conclusion.

LIMITATIONS

Limitations are the last paragraphs in the discussion. Be honest about your limitations.

How do you determine your study limitations? You're off to a good start. You've already contrasted your study with other studies in the discussion, and you have already looked for possible explanations for the contrasts. Don't forget that one possible explanation is that your study methods may have limitations. Differences between your results and other published results are the first clue to help you find the limitations of your study.

Warning: editors prefer authors who disclose all study limitations. If editors find limitations that the author didn't mention, the editors are more inclined to develop a negative feeling about the quality of the manuscript. Authors should try to state all study limitations.

BIAS

The next way to detect study limitations is to review various categories of bias. There are long lists of various types of bias, but they can also be combined in a short list: transfer bias, recording bias, reporting bias, performance bias, and selection bias.

Transfer bias is patients lost to follow-up. Journals prefer 2-year follow-up, with transfer bias of less than 20%. Transfer bias of greater than 20% should be mentioned as a limitation.

Recording bias depends on who measures the data. Patient-reported outcome forms minimize recording bias, but for objective data collection, someone other than the surgeon who performed the procedure should measure and record physical examination and other clinical outcome measures. Ideally that recorder should be blinded as to which treatment the patient received.

Reporting bias considers how outcome is reported. Outcome measures must be validated for the condition being tested or measured. To minimize recording bias, authors should select outcome measures that are commonly used in the literature. This allows study results to be compared and contrasted to other published studies, so *select the correct outcome measures before you start your study*. If you fail, your study will be limited by reporting bias, and you won't be able to compare and contrast your results with other studies, so it will be very difficult to write an interesting discussion. In other words, your conclusions may have clinical meaning, but that meaning may go unnoticed or be unappreciated.

Performance bias depends on who performs the research and who performs the surgery. Single-surgeon studies introduce bias. Multi-centered studies introduce bias. No methods can eliminate performance bias entirely, because someone always has to perform the study. In the limitations section, consider and disclose performance bias.

Selection bias occurs when 2 groups being compared have different prognoses. There is an old saying about apples and oranges, and we agree that you cannot compare apples and oranges as equals. They look different and taste different. In research, selection bias occurs when comparing 2 groups that are not equal. For example, comparing children and adults is like comparing apples and oranges.

The best way to mitigate against selection bias is prospective randomization. Another good way to minimize selection bias is to have strict study inclusion and exclusion criteria. And, such criteria must always be carefully reported. For example, a study could include children age 12 to 18 and exclude adults. These are strict inclusion and exclusion criteria that will limit the differences between patients and allow comparison of groups while minimizing selection bias.

PROSPECTIVE BY DEFINITION

Retrospective review of prospectively collected data is not prospective research.

Prospective, by definition, means the research methods, *including the research question*, are designed and written before the first patient is treated.

Warning: prospective, by definition, means that all the research methods must be written before the first patient is treated. This includes the statistical methods. Therefore, be sure to hire a statistician and write the statistical methods before you actually begin your research. Prospective research *always* lowers the chance of bias.

CONCLUSIONS

Are you feeling excited? People are always happy to reach the ending; *some even start reading the ending and don't read anything else from your paper*. So when you state the conclusion, you'll have the full attention of the reader. Don't ruin it.

The conclusion is simple. Yes or no? Do the results support your hypothesis?

Many young authors have trouble here. The conclusion must be based on the results, nothing else. But inexperienced authors always add something else. They go on and on when they shouldn't. Don't state anything that is not supported by your data. Regrettably, many people do.

The conclusion can be only one of two simple possibilities: either the hypothesis is supported by the data or it is not. Many authors forget to mention the hypothesis, either proven or disproven, in their conclusion.

Statements about future research are not appropriate for the conclusion. Such statements should be integrated within, or follow, the discussion of study limitations, just prior to the conclusion. The conclusion should not just be extended discussion. If you have to explain your conclusion, you must go back and do it in the discussion. Only then, when you have said everything that you feel like you have to say, then and only then are you finally ready to state your simple conclusion.

Warning: editors are not happy if the study conclusion is different from the conclusion in the abstract.

TITLE

Obviously the title comes before the conclusion. The problem is most authors select boring titles. Review the section on the introduction above. Controversy spices things up. After reaching the conclusion, go back and rewrite the title to make it more controversial. Remember you need to draw the reader in.

The title should be short and succinct. Put some work into it.

ABSTRACT

The abstract has 4 sections: purpose, methods, results, and conclusion.

In *Abstract:* Purpose, sum up the controversy in a single sentence, stating the purpose in terms of the hypothesis being tested (but do not actually state the hypothesis). Don't use abbreviations in the abstract.

In *Abstract:* Methods, sum up who, what, where, when, and especially why. Sum up the type of study, inclusion or exclusion criteria, primary outcome measure, consideration of statistical power, and documentation of institutional review board approval.

In *Abstract:* Results, state the *P* value, mean (range), and confidence interval.

In *Abstract:* Conclusion, remember, "Our results demonstrate . . . ," then mention limitations in terms of bias, and consider clinical relevance.

FIGURES

It is said that a picture is worth a thousand words. Include ample (but not too many) figures.

Legends

Figure Legends must "stand alone," i.e., contain a complete, take-home, educational message, as if a reader viewed only that figure without looking at any other figure or without reading the text. Be sure to point out what you want the reader to see. It may be obvious to you but the reader may miss it if you do not point it out. For anatomic or arthroscopic figures, be sure to mention patient position, side, and viewing portal. Labels are generally always helpful. The Figure Legend is equally important as the figure itself.

TABLES

Similarly, tables must "stand alone," i.e., contain a complete, take-home, educational message, as if a reader viewed only that table without looking at any other table or without reading the text. Tables should include explanatory table notes as needed.

As above, the best Results are tabulated clearly, with a brief text section pointing out the highlights of each table. To reiterate, limit textual repetitions to the Table highlights.

REFERENCES

High-impact references must be recent. Editors almost always prefer references from the last 5 years. When it comes to references, like most things in life, quality is more important than quantity. It's not a competition to have the most references.

Keep the references recent and relevant. Look for new publications and do it often during the course of your study.

ETHICS

Compliance with journal, academic, patient protection, and industry regulations is mandatory. It is incumbent upon authors to independently research these issues and insure self-regulation and compliance.

CONCLUSIONS

Choose a research question in which you are interested in knowing the answer, state your purpose and hypothesis, and prepare methods and statistical methods prospectively before treating the first patient. Be sure your conclusion is based on the results. You just put it all together and write a winning scientific paper.

James H. Lubowitz, M.D.
Gary G. Poehling, M.D.
Jón Karlsson, M.D., Ph.D.

SECTION 19

Common Mistakes in Manuscripts and How to Avoid Them

Poor quality research should not be published, award-winning research methods should be published (but sometimes require extensive revision), and average manuscripts have the best chance of being published when they are well written and all details are formatted in good order.

On the other hand, some good ideas—and even good research—never get through because of sloppy writing.

When looking at a manuscript—and now we are talking about clinical studies—there are four common mistakes. It is interesting that these mistakes are repeated over and over again in terms of writing a manuscript.

1. *The manuscript is too long*. In fact, it may be said that all manuscripts are too long. This means that manuscripts contain nonrelevant and/or very well-known facts. As an example: you start your paper on osteosynthesis of hip fractures by stating that "Hip fractures are very common . . ."

   OK, this is correct. Hip fractures are very common, but everyone in the world knows this and the paper will not be better by writing for the 5,000th time that hip fractures are common. The manuscript will just be longer and more difficult to read. Including tangentially related material is another bad way to add unnecessary length to the article. Manuscripts need to be focused. Look at your purpose and hypothesis and if something does not directly relate to these two things, then it should not be in the manuscript.
2. *Repetitions*. A good rule of thumb is that a manuscript should be as long as necessary but as short as possible. In too many manuscripts, too many issues are repeated. This is especially true for the results section versus figures and tables. We find that a well-designed table with an explanatory note is most commonly the best way to present data. Then the written results need only mention the highlights and refer the reader to the table(s).
3. *Flow*. Finally, we need to talk about the "flow" of the writing. This is very important. Your ideas must flow so your manuscript is easy to read and to follow from beginning to end. A manuscript without a logical flow will probably never be published, so you should devote a good deal of time to this. Most often, the best way to improve the flow is to make the paper shorter.

TITLE

Many times we see that the title is neutral and doesn't really say much. Instead, be direct and say something controversial. Be provocative and let people know what you mean loud and clear. They should read your work and spread the word. But, it is up to you to make them interested. Being provocative does not mean being offensive.

An example of a boring, uninteresting title versus a vivid, thought-provoking one might be: "Two-Year Results After ACL Reconstruction" versus "Major Risk of Osteoarthritis and Inferior Knee Function Two Years After ACL Reconstruction."

ABSTRACT

The important task is to make the abstract concise yet still include the purpose, the key methods, results, and conclusions. It should *not* have any back-

ground, hypothesis, or discussion. The conclusions of the abstract and the text should be the same. Generally the abstract should be no more than 300 words.

Do not include an introduction or discussion. All material should be focused on the purpose and the results. There is no need to describe the methods in great detail in the abstract. It will only make the abstract longer and destroy the flow. To destroy the flow is the worst mistake an author can make; the abstract will be more difficult to read and understand, and marginal readers will lose interest and stop reading right there.

Finally, don't let the abstract fade out into nothing. The abstract should have straightforward and clinically relevant conclusions. And importantly, if it is basic science, you must give a clear picture of the relevance to clinicians. What do the results show and what is the clinical relevance?

INTRODUCTION

This is often the most difficult part of the manuscript—how to get started? Some people never do. A good rule is, just do it—just get started. The purpose of the introduction is to frame the question that you propose to answer. This is where you the author need to draw in the reader. It should be thought-provoking and supported by the latest literature. Be careful not to slip into discussion. You do not want to explain the reasons or answer the questions in the introduction. It should be designed to just whet the appetite of the reader. Do not let it ramble or be too general, but keep it brief and focused. The last paragraph of the introduction should state clearly your purpose and then your hypothesis or what you thought you would find before you started the study. This is very important and in too many manuscripts these two essential elements are not clear at all.

A general rule is that the introduction should be no longer than one manuscript page. It must tell the readers why the study is needed and what the controversy is. Scientific controversy is good; it is not personal. Don't be afraid of it.

METHODS

Similar problems are common in the methods and results sections: too long, too vague, and do not tell the full story. Methods must be so well described that other researchers can repeat them without trouble. It should be like a cookbook. Who are you taking into your study and what are you doing to them? This is important and means that you need to have clear inclusion and exclusion criteria as well as a description of exactly how each subject was treated.

We often find that decimals are a problem. Too often authors are reporting values up to three or even four decimal places. Why is that a problem? Two reasons:

1. The flow: the reader is drowned in numbers and readability is affected.
2. More importantly, the accuracy of the measurement and methods is not mentioned in the methods section. Are your methods truly accurate to that degree? And why is the test-retest reliability measurement so infrequently reported in manuscripts? Any study is only as good as its methods and that is why the measurement's accuracy is absolutely vital.

The study cohort is often too small. This is a very common mistake. It really doesn't have anything to do with writing a manuscript, but a good explanation of why the cohort is limited is necessary and is a help to the reader. The authors may discuss such things as power, sample size calculation, and compliance, and they may mention drop-out analysis.

Statistical methods are a necessary subheading. Any study that shows equivalency requires a power analysis with an explanation of the assumptions. For basic science studies, you need to explain and provide context for the rationale of the study design.

RESULTS

The results section must be succinct. A good rule is: make it less than one manuscript page. If decimals are a problem in the methods section, they are also a problem in the results section. Often we see duplication in the results, figures, and tables. This is problematic on several levels. It destroys the flow of the manuscript and it adds to the length of the paper. The flow of the article will be enhanced when the results section is constructed to parallel the methods section.

DISCUSSION

The mistake we often see is that the discussion is too long, too general, and too vague. We suggest that you start the discussion with a sentence stating your most important findings. This should be followed by comparing and contrasting your findings with those reported in the literature. If you have done a basic science study, you need to remember that you are

writing for a clinically oriented journal. Even though you have the most wonderful study on mice and rats, you should mention the possible clinical impact.

Two important words are key factors in discussion: c*ontext* and *limitations*. Don't hide your limitations—make them visible and transparent. Classically, they should be stated at the end in the paragraph just before the conclusions. All studies have limitations; some major, some minor. And you should discuss them. Be honest about your limitations because the limitations may be the most important issue in your whole study. Why is this? A profound understanding of limitations will create new studies and new science that will lead to new understanding.

CONCLUSIONS

Conclude what you found from your data and nothing else. Too often, the conclusions section is too long and general and filled with feelings. If you have compared single- versus double-bundle ACL reconstruction, your conclusion should not be that all ACL injuries in children should be operated on because you feel that their knees will be better off after surgery. We often see trends reported in the conclusions. Trends may be in the discussion but only statistically significant findings may be in the conclusions.

REFERENCES

Two common errors that we see with references are

1. Incorrect order and incorrect format: each journal has specific instructions for references. These need to be read and carefully followed.
2. Not up to date: a possible reason for this is that the authors started the study several years back. They looked for relevant references when they started but they never updated the references by adding recent relevant citations. Why use the old ones? Bankart (1923) has been cited several thousand times. Is the manuscript really better if his study is cited once more?

Too often we see incorrect citations; what is that all about? Authors should have read the *original* publication and used that as the reference. A good example is the currently used Lysholm score that was published by Tegner and Lysholm in 1985 in *Clinical Orthopaedics and Related Research* and not by Lysholm et al. in 1982 in the *American Journal of Sports Medicine*. This error is common.

Update your references just before you send your manuscript to the editorial office. There is nothing that makes the reviewers and Editor so happy as updated references.

FIGURES AND TABLES

Figures should only be used to transmit key ideas and concepts. And don't forget that the key ideas also need to be pointed out in the figure legend. The combination of figure and legend needs to have a take-home message for the reader. Even if the point is obvious to you, it needs to be stated so the reader does not miss the point. Also, each figure/legend needs to stand on its own; the reader should not have to refer to the text to understand what you want to convey. The text should not rehash the information in the legend and vice versa. The same can be said for tables, which are the preferred method of presenting large volumes of data.

IN THE END

The good news here is that with a little care you can present your scientific work in a pleasing and accurate way that will have a high likelihood of being published.

Jón Karlsson, M.D., Ph.D.
James H. Lubowitz, M.D.
Gary G. Poehling, M.D.

REFERENCES

1. Watts G. Let's pension off the "major breakthrough." *BMJ* 2007;334:4.
2. Nordenstrom J. *Evidence-based medicine in Sherlock Holmes' footsteps.* Malden: Blackwell Publishing, 2006.
3. Bhandari M, Jain A. The need for evidence-based orthopaedics. *Indian J Orthop* 2007;41:3.
4. Claridge J, Fabian T. History and development of evidence-based medicine. *World J Surg* 2005;29:547-553.
5. Hoppe D, Bhandari M. Evidence-based orthopaedics—A brief history. *Indian J Orthop* 2008;42:104-110.
6. Packard G. Hypertrophy of one lower extremity. *Proc Am Orthop Assoc* 1889;1:27-37.
7. Cochrane A. *Effectiveness and efficiency. Random reflections on health services*. London: Nuffield Provincial Hospitals Trust, 1972.
8. Sackett D, Rosenberg W, Gray J, Haynes R, Richardson W. Evidence-based medicine: What it is and what it isn't. *BMJ* 1996;312:71-71.
9. Guyatt G. Evidence-based medicine. *ACP J Club* 1991;114: A16 (Suppl 2).
10. Evidence-Based Medicine Working Group. Evidence-based

medicine. A new approach to teaching the practice of medicine. *JAMA* 1992;268:2420-2425.
11. Covell D, Uman G, Manning P. Information needs in office practice: Are they being met? *Ann Intern Med* 1985;103:596-599.
12. Osheroff J, Forsythe D, Buchanan B, Bankowitz R, Blumenfeld B, Miller R. Physicians' information needs: Analysis of questions posed during clinical teaching. *Ann Intern Med* 1991:114:576-581.
13. Akai M. Evidence-based medicine for orthopaedic practice. *J Orthop Sci* 2002;7:731-742.
14. Hook O. Scientific communications, history, electronic journals and impact factors. *J Rehab Med* 1999;31:3-7.
15. Lindberg D, Humphreys B. Medicine and health on the Internet: The good the bad and the ugly. *JAMA* 1998;280: 1303-1304.
16. Haynes R, Sackett D. Purpose and procedure. *Evid Based Med* 1995;1:2.
17. Straus S, Richardson W, Rosenberg W, Haynes R. *Evidence-based medicine: How to practice and teach EBM*. Edinburgh: Churchill Livingstone, 2005.
18. Matzkin E, Smith E, Fornari E, Saillant J. The use of MRI for the management of suspected knee pathology by orthopaedic and nonorthopaedic practitioners. Presented at the Annual Meeting of the American Academy of Orthopaedic Surgeons, February 2011, San Diego.
19. Ebell M, Siwek J, Weiss B, et al. Strength of Recommendation Taxonomy (SORT): A patient-centered approach to grading evidence in the medical literature. *J Am Board Fam Pract* 2004;17:59-67.
20. Guyatt G, Oxman A, Vist G, et al. GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008;336:924-926.
21. Bhandari M. Evidence-based medicine: Why bother? *Arthroscopy* 2009;25:296-297.
22. Wright JG, Swiontkowski MF, Heckman JD. Introducing levels of evidence to the journal. *J Bone Joint Surg Am* 2003;85:1-3.
23. Spindler KP, Kuhn JE, Dunn W, Matthews CE, Harrell FE Jr, Dittus RS. Reading and reviewing the orthopaedic literature: A systematic, evidence-based medicine approach. *J Am Acad Orthop Surg* 2005;13:220-229.
24. Panesar SS, Philippon MJ, Bhandari M. Principles of evidence-based medicine. *Orthop Clin North Am* 2010;41:131-138.
25. Guyatt GH, Sackett DL, Sinclair JC, et al. Users' guides to the medical literature. IX. A method for grading health care recommendations. Evidence-Based Medicine Working Group. *JAMA* 1995;274:1800-1804.
26. Wright JG, Einhorn TA, Heckman JD. Grades of recommendation. *J Bone Joint Surg Am* 2005;87:1909-1910.
27. Poolman RW, Petrisor BA, Marti RK, Kerkhoffs GM, Zlowodzki M, Bhandari M. Misconceptions about practicing evidence-based orthopedic surgery. *Acta Orthop* 2007;78: 2-11.
28. Oxford Centre for Evidence Based Medicine. Available from: http://www.cebm.net/index.aspx?o=1025. Accessed November 13, 2010.
29. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. *Lancet* 2007;370:1453-1457.
30. Brozek JL, Akl EA, Alonso-Coello P, et al; GRADE Working Group. Grading quality of evidence and strength of recommendations in clinical practical guidelines. Part 1 of 3. An overview of the GRADE approach and grading quality of evidence about interventions. *Allergy* 2009;64:669-677.
31. Wright JG. A practical guide to assigning levels of evidence. *J Bone Joint Surg Am* 2007;89:1128-1130.
32. Seng K, Appleby D, Lubowitz JH. Operative versus nonoperative treatment of anterior cruciate ligament rupture in patients aged 40 years or older: An expected-value decision analysis. *Arthroscopy* 2008;24:914-920.
33. Sackett DL, Richardson WS, Rosenberg WM, Haynes RB. *Evidence-based medicine: How to teach EBM*. New York: Churchill Livingstone, 2000.
34. Ilizaliturri VM Jr, Chaidez C, Villegas P, Briseño A, Camacho-Galindo J. Prospective randomized study of 2 different techniques for endoscopic iliopsoas tendon release in the treatment of internal snapping hip syndrome. *Arthroscopy* 2009;25:159-163.
35. Niemeyer P, Köstler W, Salzmann GM, Lenz P, Kreuz PC, Südkamp NP. Autologous chondrocyte implantation for treatment of focal cartilage defects in patients age 40 years and older: A matched-pair analysis with 2-year follow-up. *Am J Sports Med* 2010;38:2410-2416.
36. Wang C, Ghalambor N, Zarins B, Warner JJ. Arthroscopic versus open Bankart repair: Analysis of patient subjective outcome and cost. *Arthroscopy* 2005;21:1219-1222.
37. Cohen M, Ferretti M, Quarteiro M, et al. Transphyseal anterior cruciate ligament reconstruction in patients with open physes. *Arthroscopy* 2009;25:831-838.
38. Bedi A, Dines J, Dines DM, et al. Use of the 70° arthroscope for improved visualization with common arthroscopic procedures. *Arthroscopy* 2010;26:1684-1696.
39. Farrokhyar F, Karanicolas PJ, Thoma A, et al. Randomized controlled trials of surgical interventions. *Ann Surg* 2010; 251:409-416.
40. Devereaux PJ, Bhandari M, Clarke M, et al. Need for expertise based randomised controlled trials. *BMJ* 2005;330:88.
41. Smith GC, Pell JP. Parachute use to prevent death and major trauma related to gravitational challenge: Systematic review of randomised controlled trials. *BMJ* 2003;327:1459-1461.
42. Solomon MJ, McLeod RS. Should we be performing more randomized controlled trials evaluating surgical operations? *Surgery* 1995;118:459-467.
43. Sauerland S, Lefering R, Neugebauer EA. The pros and cons of evidence-based surgery. *Langenbecks Arch Surg* 1999; 384:423-431.
44. Krywulak SA, Mohtadi NG, Russell ML, Sasyniuk TM. Patient satisfaction with inpatient versus outpatient reconstruction of the anterior cruciate ligament: A randomized clinical trial. *Can J Surg* 2005;48:201-206.
45. McDonald PJ, Kulkarni AV, Farrokhyar F, Bhandari M. Ethical issues in surgical research. *Can J Surg* 2010;53:133-136.
46. Mohtadi N. Arthroscopic versus open stabilization for traumatic shoulder instability. Available at: http://clinicaltrials.gov/ct2/show/NCT00251264. Accessed March 10, 2010.
47. Bednarska E, Bryant D, Devereaux PJ. Orthopaedic surgeons prefer to participate in expertise-based randomized trials. *Clin Orthop Relat Res* 2008;466:1734-1744.
48. Freedman B. Equipoise and the ethics of clinical research. *N Engl J Med* 1987;317:141-145.
49. Kitto S, Villanueva EV, Chesters J, Petrovic A, Waxman BP, Smith JA. Surgeons' attitudes towards and usage of evidence-based medicine in surgical practice: A pilot study. *ANZ J Surg* 2007;77:231-236.
50. Mohtadi N, Chan D. A RCT comparing three methods for anterior cruciate ligament reconstruction. 2010. Available at: http://clinicaltrials.gov/show/NCT00529958. Accessed March 10, 2010.
51. Mohtadi NG, Chan DS, Dainty KN, Whelan DB. Patellar tendon versus hamstring autograft for anterior cruciate ligament rupture in adults. Available at: http://www2.cochrane.org/reviews/en/protocol_C01E5E8082

E26AA201338D3DD3C94EC6.html. Accessed March 10, 2010.
52. Kitto S, Petrovic A, Gruen RL, Smith JA. Evidence-based medicine training and implementation in surgery: The role of surgical cultures. *J Eval Clin Pract*, in press.
53. Lilford R, Braunholtz D, Harris J, Gill T. Trials in surgery. *Br J Surg* 2004;91:6-16.
54. McLeod RS. Issues in surgical randomized controlled trials. *World J Surg* 1999;23:1210-1214.
55. Solomon MJ, Laxamana A, Devore L, McLeod RS. Randomized controlled trials in surgery. *Surgery* 1994;115:707-712.
56. Stirrat GM. Ethics and evidence-based surgery. *J Med Ethics* 2004;30:160-165.
57. Mohtadi NG, Hollinshead RM, Ceponis PJ, Chan DS, Fick GH. A multi-centre randomized controlled trial comparing electrothermal arthroscopic capsulorrhaphy versus open inferior capsular shift for patients with shoulder instability: Protocol implementation and interim performance: Lessons learned from conducting a multi-centre RCT [ISRCTN68224911; NCT00251160]. *Trials* 2006;7:4.
58. Farrokhyar F, Bhandari M. Practical tips for surgical research: Introduction to the series. *Can J Surg* 2010;53:67-68.
59. Thoma A, Farrokhyar F, McKnight L, Bhandari M. Practical tips for surgical research: How to optimize patient recruitment. *Can J Surg* 2010;53:205-210.
60. Farrugia P, Petrisor BA, Farrokhyar F, Bhandari M. Practical tips for surgical research: Research questions, hypotheses and objectives. *Can J Surg* 2010;53:278-281.
61. Altman DG, Schulz KF, Moher D, et al. The revised CONSORT statement for reporting randomized trials: Explanation and elaboration. *Ann Intern Med* 2001;134:663-694.
62. Moher D, Schulz KF, Altman DG. The CONSORT statement: Revised recommendations for improving the quality of reports of parallel group randomized trials. *BMC Med Res Methodol* 2001;1:2.
63. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *J Clin Epidemiol* 2010;63:834-840.
64. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials. *J Clin Epidemiol* 2010;63:e1-e37.
65. Willits K, Amendola A, Bryant D, et al. Operative versus nonoperative treatment of acute Achilles tendon ruptures: A multicenter randomized trial using accelerated functional rehabilitation. *J Bone Joint Surg Am* 2010;92:2767-2775.
66. Virk SS, Kocher MS. Adoption of new technology in sports medicine: Case studies of the Gore-Tex prosthetic ligament and of thermal capsulorrhaphy. *Arthroscopy* 2011;27:113-121.
67. Vandenbroucke JP, von Elm E, Altman DG, et al; for the STROBE Initiative. Strengthening the reporting of observational studies in epidemiology (STROBE): Explanation and elaboration. *Epidemiology* 2007;18:805-835.
68. Marsh J, Bryant D, MacDonald SJ. Older patients can accurately recall their preoperative health status six weeks following total hip arthroplasty. *J Bone Joint Surg Am* 2009;91:2827-2837.
69. Bhandari M, Joensson A. *Clinical research for surgeons.* New York: Thieme Publishing Group, 2009.
70. Vestergaard P, Rejnmark L, Mosekilde L. Fracture risk associated with use of nonsteroidal anti-inflammatory drugs, acetylsalicylic acid, and acetaminophen and the effects of rheumatoid arthritis and osteoarthritis. *Calcif Tissue Int* 2006;79:84-94.
71. Wynder EL, Graham EA. Tobacco smoking as a possible etiologic factor in bronchiogenic carcinoma; a study of 684 proved cases. *J Am Med Assoc* 1950;143:329-336.
72. Sackett DL, Wennberg JE. Choosing the best research design for each question. *BMJ* 1997;315:1636.
73. Thabane L, Thomas T, Ye C, Paul J. Posing the research question: Not so simple. *Can J Anaesth* 2009;56:71-79.
74. Marx RG, McCarty EC, Montemurno D, Altchek DW, Craig EV, Warren RF. Development of arthrosis following dislocation of the shoulder: A case-controlled study. *J Shoulder Elbow Surg* 2002;11:1-5.
75. Vandenbroucke JP, Elm EV, Altman DG, et al. Strengthening the reporting of observational studies in epidemiology (STROBE): Explanation and elaboration. *Epidemiology* 2007;18:805-835.
76. Chaudhari AM, Zelman EA, Flanigan DC, Kaeding CC, Nagaraja HN. Anterior cruciate ligament-injured subjects have smaller anterior cruciate ligaments than matched controls: A magnetic resonance imaging study. *Am J Sports Med* 2009;73:1282-1287.
77. Busse JW, Obremskey WT. Principles of designing an orthopaedic case-control study. *J Bone Joint Surg Am* 2009;91:15-20 (Suppl 3).
78. Riddle DL, Pulisic M, Pidcoe P, Johnson RE. Risk factors for plantar fasciitis: A matched case-control study. *J Bone Joint Surg Am* 2003;85:872-877.
79. Bhandari M, Morshed S, Tornetta III, P, Schemitsch EH. Design, conduct, and interpretation of nonrandomized orthopaedic studies. A practical approach. (All) evidence matters. *J Bone Joint Surg Am* 2009;91:1 (Suppl 3).
80. Lubowitz GH, Poehling GG. In defense of case series: Hip SCFE, shoulder instability and arthritis, double-bundle ACL cyclops lesions, and elbow OCD. *Arthroscopy* 2010;26:1411-1413.
81. Kooistra V, Dijkman B, Eijhorn TA, Bhandari M. How to design a good case series. *J Bone Joint Surg Am* 2009;91:21-26.
82. Hess DR. Retrospective studies and chart reviews. *Respir Care* 2004;49:1171-1174.
83. Parvizi J, Tarity TD, Conner K, Smith JB. Institutional review board approval: Why it matters. *J Bone Joint Surg Am* 2007;89:418-426.
84. McMaster WC, Sale K, Andersson G, et al. The conduct of clinical research under the HIPAA Privacy Rule. *J Bone Joint Surg Am* 2006;88:2765-2770.
85. Arendt E, Agel J, Heikes C, Griffiths H. Stress injuries to bone in college athletes. A retrospective review of experience at a single institution. *Am J Sports Med* 2003;31:959-968.
86. Poolman RW, Swiontkowski MF, Fairbank JCT, Schemitsch EH, Sprague S, de Vet HCW. Outcome instruments: Rationale for their use. *J Bone Joint Surg Am* 2009;91:41-49 (Suppl 3).
87. Suk M, Hanson BP, Norvell DC, Helfet DL. *AO handbook. Musculoskeletal outcomes measures and instruments.* New York: Thieme, 2005.
88. Portney LG, Watkins MP. *Foundations of clinical research. Applications to practice*. Ed 3. Upper Saddle River, NJ: Pearson Education, 2009.
89. Akobeng AK. Understanding the systematic reviews and meta-analysis. *Arch Dis Child* 2005;90:845-848.
90. Egger M, Smith GD. Potentials and promise. *BMJ* 1997;315:1371-1374.
91. Cook DJ, Mulrow CD, Haynes RB. Systematic reviews: Synthesis of best evidence for clinical decisions. *Ann Intern Med* 1997;126:376-380.
92. Mulrow CD. Rationale for systematic reviews. *BMJ* 1994;309:597-599.
93. Lau J, Ioannidis JP, Schmid CH. Quantitative synthesis in systematic reviews. *Ann Intern Med* 1997;127:820-826.
94. Grady D, Hearst N. Utilizing existing databases. In: Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB, eds. *Designing clinical research.* Philadelphia: Wolters Kluwer, 2007;207-211.

95. Khan KS, Kunz R, Kleijnen J, Antes G. Five steps to conducting a systematic review. *J R Soc Med* 2003;96:118-121.
96. Dickersin K, Scherer R, Lefebvre C. Identifying relevant studies for systmatic reviews. *BMJ* 1994;309:1286-1291.
97. Centre for Reviews and Dissemination. *Finding studies for systematic reviews: A resource list for researchers.* York: National Institute for Health Research, 2010;1-16.
98. Juni P, Altman GD, Mathias E. Assessing the quality of controlled clinical trials. *BMJ* 2001;323:42-46.
99. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. *BMJ* 1994;309:1351-1355.
100. Egger M, Smith GD, Phillips AN. Principles and procedures. *BMJ* 1997;315:1533-1537.
101. Cooper H, Hedges LV. *The handbook of research synthesis.* New York: Russell Sage Foundation, 1994.
102. Egger M, Smith GD. Bias in location and selection of studies. *BMJ* 1998;316:61-66.
103. Alderson P, Green S. Publication bias. The Cochrane Collaboration [homepage on the Internet]. 2002. Available from: http://www.cochrane-net.org/openlearning/html/mod15-3.htm. Accessed January 3, 2011.
104. Smith GS, Egger M, Phillips NA. Beyond the grand mean? *BMJ* 1997;315:1610-1614.
105. Bero LA, Grilli R, Grimshaw JM, Harvey E, Oxman AD, Thomson MA. Closing the gap between research and practice: An overview of systematic reviews of interventions to promote the implementation of research findings. *BMJ* 1995; 317:465-468.
106. Jadad RA, Cook DJ, Jones A, et al. Methodology and reports of systematic reviews and meta-analyses. *JAMA* 1998;280: 278-280.
107. Poolman RW, Swionkowski MF, Fairbank JCT, Schemitsch EH, Sprague S, de Vet HCW. Outcome instruments: Rational for their use. *J Bone Joint Surg Am* 2009;91:41-49.
108. Briggs KK, Kocher MS, Rodkey WG, Steadman JR. Reliability, validity, and responsiveness of the Lysholm Knee Score and Tegner Activity Scale for patients with meniscal injury of the knee. *J Bone Joint Surg Am* 2006;88:698-705.
109. Irrgang JJ, Anderson AF, Boland AL, et al. Development and validation of the International Knee Documentation Committee subjective knee form. *Am J Sports Med* 2001; 29:600-613.
110. Kane RL. Outcome measures. In: Kane R, ed. *Understanding health care outcomes research.* Gaithersburg: Aspen Publishers, 1997.
111. Karanicolas PJ, Bhandari M, Kreder H, et al. Evaluating agreement: Conducting a reliability study. *J Bone Joint Surg Am* 2009;91:99-106.
112. Portney LG, Watkins MP. *Foundations of clinical research. Applications to practice.* Ed 3. Upper Saddle River, NJ: Pearson Education, 2009.
113. Sprague S, Quigley L, Bhandari M. Survey design in orthopaedic surgery: Getting surgeons to respond. *J Bone Joint Surg Am* 2009;91:27-34.
114. Suk M, Hanson BP, Norvell DC, Helfet DL. *AO handbook. Musculoskeletal outcomes measures and instruments.* New York: Thieme, 2005.
115. Beaton DE. Understanding the relevance of measured change through studies of responsiveness. *Spine* 2000;25: 3192-3199.
116. Ware JE Jr, Kosinski M, Keller SD. A 12-item Short-Form health survey: Construction of scales and preliminary test of reliability and validity. *Med Care* 1996;34:220-233.
117. Shoukri MM, Asyali MH, Donner A. Sample size requirements for the design of reliability study: Review and new results. *Stat Methods Med Res* 2004;13:251-271.
118. Thoma A, Sprague S, Temple C, Archibald S. The role of the randomized controlled trial in plastic surgery. *Clin Plast Surg* 2008;35:275-284.
119. Wright JG, Young NL. The patient-specific index: Asking patients what they want. *J Bone Joint Surg Am* 1997;79:974-983.
120. Jackowski D, Guyatt GH. A guide to health measurement. *Clin Orthop Relat Res* 2003:80-89.
121. Kirkley A, Griffin S, McLintock H, Ng L. The development and evaluation of a disease-specific quality of life measurement tool for shoulder instability: The Western Ontario Shoulder Instability Index (WOSI). *Am J Sports Med* 1998; 26:764-772.
122. Angst F, Goldhahn J, Pap G, et al. Cross-cultural adaptation, reliability and validity of the German Shoulder Pain and Disability Index (SPADI). *Rheumatology* 2007;46:87-92.
123. Bot SD, Terwee CB, van der Windt DA, Bouter LM, Dekker J, de Vet HCW. Clinimetric evaluation of shoulder disability questionnaires: A systematic review of the literature. *Ann Rheum Dis* 2004;63:335-341.
124. Marx RG, Jones EC, Allen AA. Reliability, validity, and responsiveness of four knee outcome scales for athletic patients. *J Bone Joint Surg Am* 2001;83:1459-1469.
125. Dias JJ, Bhowal B, Wildin CJ, Thompson JR. Assessing the outcome of disorders of the hand. Is the patient evaluation measure reliable, valid, responsive and without bias? *J Bone Joint Surg Br* 2001;83:235-240.
126. World Health Organization. *Preamble to the Constitution of the World Health Organization as adopted by the International Health Conference.* No. 2. Geneva: World Health Organization, 1948.
127. Hale SA, Hertel J. Reliability and sensitivity of the Foot and Ankle Disability Index in subjects with chronic ankle instability. *J Athl Train* 2005;40:35-40.
128. Irrgang JJ, Snyder-Mackler L, Wainner RS, Fu FH, Harner CD. Development of a patient-reported measure of function of the knee. *J Bone Joint Surg Am* 1998;80:1132-1145.
129. Saltzman CL, Domsic RT, Baumhauer JF, et al. Foot and ankle research priority: Report from the Research Council of the American Orthopaedic Foot and Ankle Society. *Foot Ankle Int* 1997;18:447-448.
130. Martin RL, Burdett RG, Irrgang JJ. Development of the Foot and Ankle Disability Index (FADI). *J Orthop Sports Phys Ther* 1999;29:A32-A33.
131. Alderman AK, Chung KC. Measuring outcomes in hand surgery. *Clin Plast Surg* 2008;35:239-250.
132. Marx RG. Knee rating scales. *Arthroscopy* 2003;19:1103-1108.
133. Streiner DL, Norman GR. *Health measurements scales: A practical guide to their development and use.* Oxford: Oxford University Press, 1989.
134. Marx RG, Menezes A, Horovitz L, Jones EC, Warren RF. A comparison of two time intervals for test-retest reliability of health status instruments. *J Clin Epidemiol* 2003;56:730-735.
135. Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. *Ann Intern Med* 1993;118:622-629.
136. Vangsness CT Jr, Mac P, Requa R, Garrick J. Review of outcome instruments for evaluation of anterior cruciate ligament reconstruction. *Bull Hosp Jt Dis* 1995;54:25-29.
137. Kirkley A, Griffin S, Dainty K. Scoring systems for the functional assessment of the shoulder. *Arthroscopy* 2003; 19:1109-1120.
138. Rowe CR, Patel D, Southmayd WW. The Bankart procedure: A long-term end-result study. *J Bone Joint Surg Am* 1978; 60:1-16.
139. Amstutz HC, Sew Hoy AL, Clarke IC. UCLA anatomic total shoulder arthroplasty. *Clin Orthop Relat Res* 1981;155:7-20.
140. Ellman H, Hanker G, Bayer M. Repair of the rotator cuff. End-result study of factors influencing reconstruction. *J Bone Joint Surg Am* 1986;68:1136-1144.
141. Romeo AA, Bach BR Jr, O'Halloran KL. Scoring systems for shoulder conditions. *Am J Sports Med* 1996;24:472-476.

142. Constant CR, Murley AH. A clinical method of functional assessment of the shoulder. *Clin Orthop Relat Res* 1987;214: 160-164.
143. Conboy VB, Morris RW, Kiss J, Carr AJ. An evaluation of the Constant-Murley shoulder assessment. *J Bone Joint Surg Br* 1996;78:229-232.
144. Richards RR, An K-N, Bigliani LU, et al. A standardized method for the assessment of shoulder function. *J Shoulder Elbow Surg* 1994;3:347-352.
145. L'Insalata JC, Warren RF, Cohen SB, Altchek DW, Peterson MG. A self-administered questionnaire for assessment of symptoms and function of the shoulder. *J Bone Joint Surg Am* 1997;79:738-748.
146. Hudak PL, Amadio PC, Bombardier C. Development of an upper extremity outcome measure: The DASH (disabilities of the arm, shoulder and hand) [corrected]. The Upper Extremity Collaborative Group (UECG). *Am J Ind Med* 1996;29: 602-608.
147. Solaway S, Beaton DE, McConnel S, Bombardier C. *The DASH outcome measure user's manual.* Toronto: Institute for Work & Health, 2002.
148. Marx RG, Hogg-Johnson S, Hudak P, et al. A comparison of patients' responses about their disability with and without attribution to their affected area. *J Clin Epidemiol* 2001;54: 580-586.
149. Kirkley A, Alvarez C, Griffin S. The development and evaluation of a disease-specific quality-of-life questionnaire for disorders of the rotator cuff: The Western Ontario Rotator Cuff Index. *Clin J Sport Med* 2003;13:84-92.
150. Kirkley A, Griffin S, McLintock H, Ng L. The development and evaluation of a disease-specific quality of life measurement tool for shoulder instability. The Western Ontario Shoulder Instability Index (WOSI). *Am J Sports Med* 1998; 26:764-772.
151. Lo IK, Griffin S, Kirkley A. The development of a disease-specific quality-of-life measurement tool for osteoarthritis of the shoulder: The Western Ontario Osteoarthritis of the Shoulder (WOOS) index. *Osteoarthritis Cartilage* 2001;9: 771-778.
152. Dawson J, Fitzpatrick R, Carr A. Questionnaire on the perceptions of patients about shoulder surgery. *J Bone Joint Surg Br* 1996;78:593-600.
153. Dawson J, Fitzpatrick R, Carr A. The assessment of shoulder instability. The development and validation of a questionnaire. *J Bone Joint Surg Br* 1999;81:420-426.
154. Dawson J, Hill G, Fitzpatrick R, Carr A. The benefits of using patient-based methods of assessment. Medium-term results of an observational study of shoulder surgery. *J Bone Joint Surg Br* 2001;83:877-882.
155. Hollinshead RM, Mohtadi NG, Vande Guchte RA, Wadey VM. Two 6-year follow-up studies of large and massive rotator cuff tears: Comparison of outcome measures. *J Shoulder Elbow Surg* 2000;9:373-381.
156. Guyatt GH, Townsend M, Berman LB, Keller JL. A comparison of Likert and visual analogue scales for measuring change in function. *J Chronic Dis* 1987;40:1129-1133.
157. Kirshner B, Guyatt G. A methodological framework for assessing health indices. *J Chronic Dis* 1985;38:27-36.
158. Eastlack ME, Axe MJ, Snyder-Mackler L. Laxity, instability, and functional outcome after ACL injury: Copers versus noncopers. *Med Sci Sports Exerc* 1999;31:210-215.
159. Heckman JD. Are validated questionnaires valid? *J Bone Joint Surg Am* 2006;88:446.
160. Neeb TB, Aufdemkampe G, Wagener JH, Mastenbroek L. Assessing anterior cruciate ligament injuries: The association and differential value of questionnaires, clinical tests, and functional tests. *J Orthop Sports Phys Ther* 1997;26:324-331.
161. Risberg MA, Holm I, Steen H, Beynnon BD. Sensitivity to changes over time for the IKDC form, the Lysholm score, and the Cincinnati knee score. A prospective study of 120 ACL reconstructed patients with a 2-year follow-up. *Knee Surg Sports Traumatol Arthrosc* 1999;7:152-159.
162. Sernert N, Kartus J, Kohler K, et al. Analysis of subjective, objective and functional examination tests after anterior cruciate ligament reconstruction. A follow-up of 527 patients. *Knee Surg Sports Traumatol Arthrosc* 1999;7:160-165.
163. Kocher MS, Steadman JR, Briggs K, Zurakowski D, Sterett WI, Hawkins RJ. Determinants of patient satisfaction with outcome after anterior cruciate ligament reconstruction. *J Bone Joint Surg Am* 2002;84:1560-1572.
164. Marx RG, Stump TJ, Jones EC, Wickiewicz TL, Warren RF. Development and evaluation of an activity rating scale for disorders of the knee. *Am J Sports Med* 2001;29:213-218.
165. Bollen S, Seedhom BB. A comparison of the Lysholm and Cincinnati knee scoring questionnaires. *Am J Sports Med* 1991;19:189-190.
166. Sgaglione NA, Del Pizzo W, Fox JM, Friedman MJ. Critical analysis of knee ligament rating systems. *Am J Sports Med* 1995;23:660-667.
167. Barber-Westin SD, Noyes FR, McCloskey JW. Rigorous statistical reliability, validity, and responsiveness testing of the Cincinnati knee rating system in 350 subjects with uninjured, injured, or anterior cruciate ligament-reconstructed knees. *Am J Sports Med* 1999;27:402-416.
168. Marx RG, Jones EC, Allen AA, et al. Reliability, validity, and responsiveness of four knee outcome scales for athletic patients. *J Bone Joint Surg Am* 2001;83:1459-1469.
169. Hefti F, Muller W. Current state of evaluation of knee ligament lesions. The new IKDC knee evaluation form. *Orthopade* 1993;22:351-362 (in German).
170. Irrgang JJ, Anderson AF. Development and validation of health-related quality of life measures for the knee. *Clin Orthop Relat Res* 2002;402:95-109.
171. Irrgang JJ, Anderson AF, Boland AL, et al. Responsiveness of the International Knee Documentation Committee Subjective Knee Form. *Am J Sports Med* 2006;34:1567-1573.
172. Lysholm J, Gillquist J. Evaluation of knee ligament surgery results with special emphasis on use of a scoring scale. *Am J Sports Med* 1982;10:150-154.
173. Tegner Y, Lysholm J. Rating systems in the evaluation of knee ligament injuries. *Clin Orthop Relat Res* 1985;198: 43-49.
174. Hoher J, Munster A, Klein J, Eypasch E, Tiling T. Validation and application of a subjective knee questionnaire. *Knee Surg Sports Traumatol Arthrosc* 1995;3:26-33.
175. Lukianov AV, Gillquist J, Grana WA, DeHaven KE. An anterior cruciate ligament (ACL) evaluation format for assessment of artificial or autologous anterior cruciate reconstruction results. *Clin Orthop Relat Res* 1987;218:167-180.
176. Irrgang JJ, Ho H, Harner CD, Fu FH. Use of the International Knee Documentation Committee guidelines to assess outcome following anterior cruciate ligament reconstruction. *Knee Surg Sports Traumatol Arthrosc* 1998;6:107-114.
177. Bengtsson J, Mollborg J, Werner S. A study for testing the sensitivity and reliability of the Lysholm knee scoring scale. *Knee Surg Sports Traumatol Arthrosc* 1996;4:27-31.
178. Wright RW. Knee injury outcomes measures. *J Am Acad Orthop Surg* 2009;17:31-39.
179. Williams GN, Taylor DC, Gangel TJ, Uhorchak JM, Arciero RA. Comparison of the single assessment numeric evaluation method and the Lysholm score. *Clin Orthop Relat Res* 2000; 373:184-192.
180. Roos EM, Roos HP, Lohmander LS, Ekdahl C, Beynnon BD. Knee Injury and Osteoarthritis Outcome Score (KOOS)—Development of a self-administered outcome measure. *J Orthop Sports Phys Ther* 1998;28:88-96.
181. Roos EM, Ostenberg A, Roos H, Ekdahl C, Lohmander LS. Long-term outcome of meniscectomy: Symptoms, function,

and performance tests in patients with or without radiographic osteoarthritis compared to matched controls. *Osteoarthritis Cartilage* 2001;9:316-324.
182. Roos EM, Roos HP, Ryd L, Lohmander LS. Substantial disability 3 months after arthroscopic partial meniscectomy: A prospective study of patient-relevant outcomes. *Arthroscopy* 2000;16:619-626.
183. W-Dahl A, Toksvig-Larsen S, Roos EM. A 2-year prospective study of patient-relevant outcomes in patients operated on for knee osteoarthritis with tibial osteotomy. *BMC Musculoskelet Disord* 2005;6:18.
184. Mohtadi N. Development and validation of the quality of life outcome measure (questionnaire) for chronic anterior cruciate ligament deficiency. *Am J Sports Med* 1998;26:350-359.
185. Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt LW. Validation study of WOMAC: A health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. *J Rheumatol* 1988;15:1833-1840.
186. Bellamy N. Outcome measurement in osteoarthritis clinical trials. *J Rheumatol Suppl* 1995;43:49-51.
187. Dawson J, Fitzpatrick R, Murray D, Carr A. Questionnaire on the perceptions of patients about total knee replacement. *J Bone Joint Surg Br* 1998;80:63-69.
188. Mann G, Nyska M, Hetsroni I, Karlsson J. Scoring systems for evaluating ankle function. *Foot Ankle Clin* 2006;11:509-519.
189. Zengerink M, Struijs PA, Tol JL, van Dijk CN. Treatment of osteochondral lesions of the talus: A systematic review. *Knee Surg Sports Traumatol Arthrosc* 2010;18:238-246.
190. Junge A, Langevoort G, Pipe A, et al. Injuries in team sport tournaments during the 2004 Olympic Games. *Am J Sports Med* 2006;34:565-576.
191. Good CJ, Jones MA, Lingstone BN. Reconstruction of the lateral ligament of the ankle. *Injury* 1975;7:63-65.
192. Sefton GK, George J, Fitton JM, McMullen H. Reconstruction of the anterior talofibular ligament for the treatment of the unstable ankle. *J Bone Joint Surg Br* 1979;61:352-354.
193. St Pierre R, Allman F Jr, Bassett FH III, Goldner JL, Fleming LL. A review of lateral ankle ligamentous reconstructions. *Foot Ankle* 1982;3:114-123.
194. Karlsson J, Peterson L. Evaluation of ankle joint function: The use of a scoring scale. *Foot Ankle Int* 1991;1:15-19.
195. Kaikkonen A, Kannus P, Jarvinen M. A performance test protocol and scoring scale for the evaluation of ankle injuries. *Am J Sports Med* 1994;22:462-469.
196. Kitaoka HB, Alexander IJ, Adelaar RS, Nunley JA, Myerson MS, Sanders M. Clinical rating systems for the ankle-hindfoot, midfoot, hallux, and lesser toes. *Foot Ankle Int* 1994;15:349-353.
197. de Bie RA, de Vet HC, van den Wildenberg FA, Lenssen T, Knipschild PG. The prognosis of ankle sprains. *Int J Sports Med* 1997;18:285-289.
198. Parker J, Nester CJ, Long AF, Barrie J. The problem with measuring patient perceptions of outcome with existing outcome measures in foot and ankle surgery. *Foot Ankle Int* 2003;24:56-60.
199. Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating patient-based outcome measures for use in clinical trials. *Health Technol Assess* 1998;2:1-74.
200. Rozzi SL, Lephart SM, Sterner R, Kuligowski L. Balance training for persons with functionally unstable ankles. *J Orthop Sports Phys Ther* 1999;29:478-486.
201. Roos EM, Brandsson S, Karlsson J. Validation of the foot and ankle outcome score for ankle ligament reconstruction. *Foot Ankle Int* 2001;22:788-794.
202. Hale SA, Hertel J. Reliability and sensitivity of the foot and ankle disability index in subjects with chronic ankle instability. *J Athl Train* 2005;40:35-40.
203. Martin RL, Irrgang JJ, Burdett RG, Conti SF, Van Swearingen JM. Evidence of validity for the Foot and Ankle Ability Measure (FAAM). *Foot Ankle Int* 2005;26:968-983.
204. Eechaute C, Vaes P, Van Aerschot L, Asman S, Duquet W. The clinimetric qualities of patient-assessed instruments for measuring chronic ankle instability: A systematic review. *BMC Musculoskelet Disord* 2007;8:6.
205. Dobbs HS. Survivorship of total hip replacements. *J Bone Joint Surg Br* 1980;62:168-173.
206. Murray DW. Survival analysis. In: Pynsent PB, Fairbank JCT, Carr AJ, eds. *Assessment methodology in orthopaedics*. Oxford: Reed Educational and Professional Publishing, 1997;19-28.
207. Fennema P, Lubsen J. Survival analysis in total joint replacement. *J Bone Joint Surg Br* 2010;92:701-706.
208. Murray DW, Carr AJ, Bulstrode C. Survival analysis of joint replacements. *J Bone Joint Surg Br* 1993;75:697-704.
209. Rosenberg N, Neumann L Modi A, Mersich IJ, Wallace AW. Improvements in survival of the uncemented Nottingham total shoulder prosthesis: A prospective comparative study. *BMC Musculoskeletal Disorders* 2007;8:76.
210. Carr AJ, Morris RW, Murray DW, Pynsent PB. Survival analysis in joint replacement surgery. *J Bone Joint Surg Br* 1993;75:178-182.
211. Ferdinand RD, Pinder IM. Survival analysis of joint replacements. *J Bone Joint Surg Br* 1997;79:878.
212. Rothman KJ. Estimation of confidence limits for the cumulative probability of survival in life table analysis. *J Chronic Dis* 1978;31:557-560.
213. Dawson-Saunders B, Trapp RG. *Basic & clinical biostatistics*. Ed 2. Norwalk, CT: Appleton & Lange, 1994;200-201.
214. *Mars Climate Orbiter Mishap Investigation Board Phase I Report.* 1999. Available at: ftp://ftp.hq.nasa.gov/pub/pao/reports/1999/MCO_report.pdf. Accessed March 8, 2011.
215. Kraemer HC. Pitfalls of multisite randomized clinical trials of efficacy and effectiveness. *Schizophr Bull* 2000;26:533-541.
216. Bhandari M, Tornetta P III, Ellis T, et al. Hierarchy of evidence: Differences in results between non-randomized studies and randomized trials in patients with femoral neck fractures. *Arch Orthop Trauma Surg* 2004;124:10-16.
217. Dijkman BG, Kooistra BW, Pemberton J, Sprague S, Hanson BP, Bhandari M. Can orthopedic trials change practice? *Acta Orthop* 2010;81:122-125.
218. Hayes K, Walton JR, Szomor ZR, Murrell GA. Reliability of five methods for assessing shoulder range of motion. *Aust J Physiother* 2001;47:289-294.
219. van Trijffel E, van de Pol RJ, Oostendorp RA, Lucas C. Inter-rater reliability for measurement of passive physiological movements in lower extremity joints is generally low: A systematic review. *J Physiother* 2010;56:223-235.
220. van de Pol RJ, van Trijffel E, Lucas C. Inter-rater reliability for measurement of passive physiological range of motion of upper extremity joints is better if instruments are used: A systematic review. *J Physiother* 2010;56:7-17.
221. Jordan K, Dziedzic K, Jones PW, Ong BN, Dawes PT. The reliability of the three-dimensional FASTRAK measurement system in measuring cervical spine and shoulder range of motion in healthy subjects. *Rheumatology (Oxford)* 2000;39:382-388.
222. Gerhardt JJ, Rondinelli RD. Goniometric techniques for range-of-motion assessment. *Phys Med Rehabil Clin N Am* 2001;12:507-527.
223. Terwee CB, Mokkink LB, Steultjens MP, Dekker J. Performance-based methods for measuring the physical function of patients with osteoarthritis of the hip or knee: A systematic review of measurement properties. *Rheumatology (Oxford)* 2006;45:890-902.
224. Impellizzeri FM, Marcora SM. Test validation in sport phys-

iology: Lessons learned from clinimetrics. *Int J Sports Physiol Perform* 2009;4:269-277.
225. Robinson ME, Dannecker EA. Critical issues in the use of muscle testing for the determination of sincerity of effort. *Clin J Pain* 2004;20:392-398.
226. Constant CR. Schulterfunktionsbeurteilung. *Orthopade* 1991; 20:289-294.
227. Widler KS, Glatthorn JF, Bizzini M, et al. Assessment of hip abductor muscle strength. A validity and reliability study. *J Bone Joint Surg Am* 2009;91:2666-2672.
228. Thomason K, Smith KL. The reliability of measurements taken from computer-stored digitalised x-rays of acute distal radius fractures. *J Hand Surg Eur Vol* 2008;33:369-372.
229. Bensoussan L, Viton JM, Barotsis N, Delarque A. Evaluation of patients with gait abnormalities in physical and rehabilitation medicine settings. *J Rehabil Med* 2008;40:497-507.
230. van der Leeden M, Steultjens MP, Terwee CB, et al. A systematic review of instruments measuring foot function, foot pain, and foot-related disability in patients with rheumatoid arthritis. *Arthritis Rheum* 2008;59:1257-1269.
231. Lane NE, Nevitt MC, Genant HK, Hochberg MC. Reliability of new indices of radiographic osteoarthritis of the hand and hip and lumbar disc degeneration. *J Rheumatol* 1993;20: 1911-1918.
232. Mast NH, Impellizzeri F, Keller S, Leunig M. Reliability and agreement of measures used in radiographic evaluation of the adult hip. *Clin Orthop Relat Res* 2011;469:188-199.
233. Martin J, Marsh JL, Nepola JV, Dirschl DR, Hurwitz S, DeCoster TA. Radiographic fracture assessments: Which ones can we reliably make? *J Orthop Trauma* 2000;14:379-385.
234. Karanicolas PJ, Bhandari M, Walter SD, et al. Interobserver reliability of classification systems to rate the quality of femoral neck fracture reduction. *J Orthop Trauma* 2009;23: 408-412.
235. Corrales LA, Morshed S, Bhandari M, Miclau T III. Variability in the assessment of fracture-healing in orthopaedic trauma studies. *J Bone Joint Surg Am* 2008;90:1862-1868.
236. Blokhuis TJ, de Bruine JH, Bramer JA, et al. The reliability of plain radiography in experimental fracture healing. *Skeletal Radiol* 2001;30:151-156.
237. Bhandari M, Guyatt G, Tornetta P III, et al. Study to prospectively evaluate reamed intramedually nails in patients with tibial fractures (S.P.R.I.N.T.): Study rationale and design. *BMC Musculoskelet Disord* 2008;9:91.
238. Goldhahn S, Kralinger F, Rikli D, Marent M, Goldhahn J. Does osteoporosis increase complication risk in surgical fracture treatment? A protocol combining new endpoints for two prospective multicentre open cohort studies. *BMC Musculoskelet Disord* 2010;11:256.
239. Andersen T, Christensen FB, Langdahl BL, et al. Fusion mass bone quality after uninstrumented spinal fusion in older patients. *Eur Spine J* 2010;19:2200-2208.
240. Gallinaro P, Masse A, Leonardi F, Buratti CA, Boggio F, Piana R. Eight- to ten-year results of a variable geometry stem. *Orthopedics* 2007;30:954-958.
241. Glassman AH, Bobyn JD, Tanzer M. New femoral designs: Do they influence stress shielding? *Clin Orthop Relat Res* 2006;453:64-74.
242. Gruen TA, McNeice GM, Amstutz HC. "Modes of failure" of cemented stem-type femoral components: A radiographic analysis of loosening. *Clin Orthop Relat Res* 1979:17-27.
243. Blake GM. Replacing DXA scanners: Cross-calibration with phantoms may be misleading. *Calcif Tissue Int* 1996;59:1-5.
244. Pearson J, Ruegsegger P, Dequeker J, et al. European semi-anthropomorphic phantom for the cross-calibration of peripheral bone densitometers: Assessment of precision accuracy and stability. *Bone Miner* 1994;27:109-120.
245. Jette A. Functional disability and rehabilitation of the aged. *Top Geriatr Rehabil* 1986;1:1-7.
246. Guyatt GH, Rennie D. *Users' guides to the medical literature: Essentials of evidence-based clinical practice.* Chicago: American Medical Association, 2002.
247. Bhandari M, Guyatt G, Montori V. User's guide to the orthopaedic literature: How to use a systematic literature review. *J Bone Joint Surg Am* 2002;84:1672-1682.
248. Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. *Ann Intern Med* 1993;118:622-629.
249. Verbrugge LM, Jette AM. The disablement process. *Soc Sci Med* 1994;38:1-14.
250. Lezzoni LI, Greenberg MS. Capturing and classifying functional status information in administrative databases. *Health Care Finance Rev* 2003;24:61-76.
251. Scientific Advisory Committee of the Medical Outcomes Trust. Assessing health status and quality-of-life instruments: Attributes and review criteria. *Qual Life Res* 2002;11:193-205.
252. Lodhia P, Slobogean GP, Noonen VK, et al. Patient-reported outcome instruments for femoracetabular impingement and hip labral pathology: A systematic review of the clinimetric evidence. *Arthroscopy* 2011;27:279-286.
253. Hoang-Kim A, Bhandari M, Beaton D, Kulkarni A, Schemitsch E. Functional status and disability tools in hip fracture RCTs are not pragmatic enough. P265. Presented at the American Academy of Orthopaedic Surgeons Annual Meeting, New Orleans, LA, March 9-13, 2010.
254. Brena SF, Sanders SH, Motoyama H. American and Japanese chronic low back pain patients: Cross-cultural similarities and differences. *Clin J Pain* 1990;6:118-124.
255. Guillemin F, Bombardier C, Beaton D. Cross-cultural adaptation of health-related quality of life measures: Literature review and proposed guidelines [see comments]. *J Clin Epidemiol* 1993;46:1417-1432.
256. Beaton D, Bombardier C, Guillemin F, Bosi Ferraz M. *Recommendations for the cross cultural adaptation of health status measures*. Rosemont, IL: American Academy of Orthopaedic Surgeons, 2002.
257. Goldstein FC, Strasser DC, Woodard JL, et al. Functional outcome of cognitively impaired hip fracture patients on a geriatric rehabilitation unit. *J Am Geriatr Soc* 1997;45:35-42.
258. Vellinga A. To know or not to be: Development of an instrument to assess decision-making capacity of cognitively impaired elderly patients. Amsterdam: Vrije University, 2006 (PhD thesis).
259. Hoang-Kim A, Beaton D, Bhandari M, Schemitsch E. HRQOL measures are underutilized in hip fracture patients with severe cognitive impairment. P263. Presented at the American Academy of Orthopaedic Surgeons Annual Meeting, New Orleans, LA, March 9-13, 2010.
260. Van Spall HG, Toren A, Kiss A, Fowler RA. Eligibility criteria of randomized controlled trials published in high-impact general medical journals: A systematic sampling review. *JAMA* 2007;297:1233-1240.
261. *ICH harmonised tripartite guideline for statistical principles for clinical trials*. Richmond, England: Brookwood Medical Publications, 1998.
262. Goldhahn S, Sawaguchi T, Audige L, et al. Complication reporting in orthopaedic trials. A systematic review of randomized controlled trials. *J Bone Joint Surg Am* 2009;91: 1847-1853.
263. Hutchinson D, ed. *The trial investigator's GCP handbook: A practical guide to ICH requirements*. Richmond, England: Brookwood Medical Publications, 1997.
264. Carlesso LC, MacDermid JC, Santaguida LP. Standardization of adverse event terminology and reporting in orthopaedic

physical therapy: Application to the cervical spine. *J Orthop Sports Phys Ther* 2010;40:455-463.
265. Corrales LA, Morshed S, Bhandari M, Miclau T III. Variability in the assessment of fracture-healing in orthopaedic trauma studies. *J Bone Joint Surg Am* 2008;90:1862-1868.
266. Davis BJ, Roberts PJ, Moorcroft CI, Brown MF, Thomas PB, Wade RH. Reliability of radiographs in defining union of internally fixed fractures. *Injury* 2004;35:557-561.
267. Morshed S, Corrales L, Genant H, Miclau T III. Outcome assessment in clinical trials of fracture-healing. *J Bone Joint Surg Am* 2008;90:62-67 (Suppl 1).
268. International Organization for Standardization. ISO_14155-1. Clinical investigation of medical devices for human subjects. Part 1: General requirements. Geneva: International Organization for Standardization, 2003;22
269. USFDA. Guidance for clinical trial sponsors. Establishment and operation of clinical trial data monitoring committees. *CBER* 2006:1-34.

AUTHORS

Laurent Audigé, D.V.M., Ph.D., AO Clinical Investigation and Documentation, AO Foundation, Dübendorf, Switzerland.
Olufemi R. Ayeni, M.D., F.R.C.S.C., Department of Surgery, McMaster University, Hamilton, Ontario, Canada.
Mohit Bhandari, M.D., Ph.D., F.R.C.S.C., Division of Orthopaedic Surgery, Department of Surgery, McMaster University, Hamilton, Ontario, Canada.
Brian W. Boyle, B.A., Hospital for Special Surgery, New York, New York, U.S.A.
Karen K. Briggs, M.P.H., Steadman Philippon Research Institute, Vail, Colorado, U.S.A.
Kevin Chan, M.D., Division of Orthopaedic Surgery, Department of Surgery, McMaster University, Hamilton, Ontario, Canada.
Kira Chaney-Barclay, M.P.H., Steadman Philippon Research Institute, Vail, Colorado, U.S.A.
Huong T. Do, M.A., Epidemiology & Biostatistics Core, Hospital for Special Surgery, New York, New York, U.S.A.
Mario Ferretti, M.D., Ph.D., Department of Orthopaedic Surgery and Traumatology, Orthopaedic Sports Medicine Division, Universidade Federal de Sao Paulo, São Paulo, Brazil.
Freddie H. Fu, M.D., Department of Orthopaedic Surgery, University of Pittsburgh, Pittsburgh, Pennsylvania, U.S.A.
Jörg Goldhahn, M.D., M.A.S., Institute for Biomechanics, ETH Zurich and Schulthess Clinic, Zurich, Switzerland.
Sabine Goldhahn, M.D., AO Clinical Investigation and Documentation, AO Foundation, Dübendorf, Switzerland.
Chisa Hidaka, M.D., Epidemiology & Biostatistics Core, Hospital for Special Surgery, New York, New York, U.S.A.
Amy Hoang-Kim, M.Sc., Ph.D.(cand), Department of Medical Sciences, University of Toronto, St. Michael's Hospital, Toronto, Ontario, Canada.
Jón Karlsson, M.D., Ph.D., Department of Orthopaedic Surgery, Sahlgrenska University Hospital/Mölndal, Mölndal, Sweden.
Aaron J. Krych, M.D., Department of Orthopaedic Surgery, Weill Cornell Medical College, Hospital for Special Surgery, New York, New York, U.S.A.
Robert F. LaPrade, M.D., Ph.D., Steadman Philippon Research Institute, Vail, Colorado, U.S.A.
Bruce A. Levy, M.D., Department of Orthopaedic Surgery, Mayo Clinic, Rochester, Minnesota, U.S.A.
James H. Lubowitz, M.D., Taos Orthopaedic Institute, Taos, New Mexico, U.S.A.
Stephen Lyman, Ph.D., Epidemiology & Biostatistics Core, Hospital for Special Surgery, New York, New York, U.S.A.
Yan Ma, Ph.D., Epidemiology & Biostatistics Core, Hospital for Special Surgery, New York, New York, U.S.A.
Robert G. Marx, M.D., M.Sc., F.R.C.S.C., Hospital for Special Surgery, New York, New York, U.S.A.
Nicholas Mohtadi, M.D., M.Sc., F.R.C.S.C., University of Calgary Sport Medicine Centre, Calgary, Alberta, Canada.
Giulio Maria Marcheggiani Muccioli, M.D., Laboratorio di Biomeccanica ed Innovazione Tecnologica, Istituto Ortopedico Rizzoli, University of Bologna, Bologna, Italy.
Norimasa Nakamura, M.D., Ph.D., Institution for Medical Science in Sports, Osaka Health Science University, Osaka City, Osaka, Japan.
Joseph Nguyen, M.P.H., Epidemiology & Biostatistics Core, Hospital for Special Surgery, New York, New York, U.S.A.

Gary G. Poehling, M.D., Department of Orthopaedics, Wake Forest University Baptist Medical Center, Winston-Salem, North Carolina, U.S.A.
Lauren E. Roberts, M.Sc., Department of Orthopaedic Surgery, McMaster University, Hamilton, Ontario, Canada.
Nahum Rosenberg, M.D., Department of Orthopaedic Surgery, Rambam Medical Center and Ruth and Bruce Rappaport Faculty of Medicine, Technion–Israel Institute of Technology, Haifa, Israel.
Kevin P. Shea, M.D., Department of Orthopaedic Surgery, University of Connecticut Health Center, Farmington, Connecticut, U.S.A.
Zahra N. Sohani, M.Sc., Department of Orthopaedics, McMaster University, Hamilton, Ontario, Canada.
Michael Soudry, M.D., Department of Orthopaedic Surgery, Rambam Medical Center and Ruth and Bruce Rappaport Faculty of Medicine, Technion–Israel Institute of Technology, Haifa, Israel.
Sophocles Voineskos, M.D., Department of Surgery, McMaster University, Hamilton, Ontario, Canada.
Stefano Zaffagnini, M.D., Laboratorio di Biomeccanica ed Innovazione Tecnologica, Istituto Ortopedico Rizzoli, University of Bologna, Bologna, Italy.