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. Rating 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.

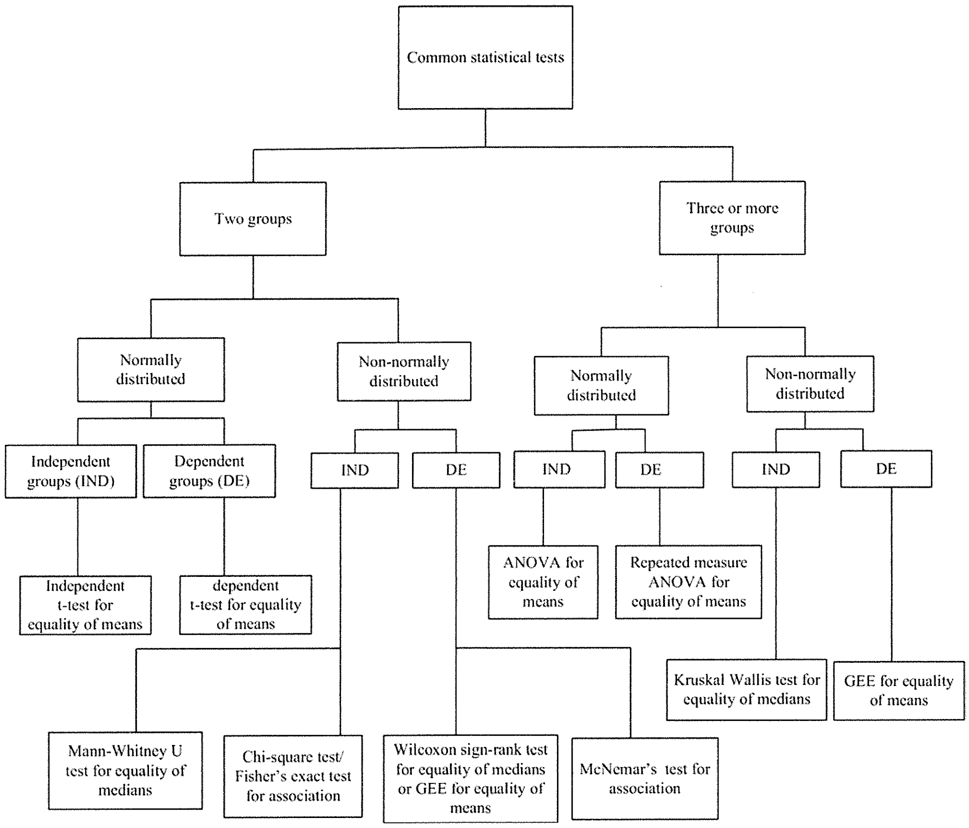
Common statistical tests
Two groups
Three or more groups
Normally distributed
Non-normally distributed
Normally distributed
Non-normally distributed
Independent groups (IND)
Dependent groups (DE)
IND
DE
IND
DE
IND
DE
Independent t-test for equality of means
dependent t-test for equality of means
ANOVA for equality of means
Repeated measure ANOVA for equality of means
Kruskal Wallis test for equality of medians
GEE for equality of means
Mann-Whitney U test for equality of medians
Chi-square test/ Fisher's exact test for association
Wilcoxon sign-rank test for equality of medians or GEE for equality of means
McNemar's test for association

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_0 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 $\mu = \mu_0$ 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 $\neq$ 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