Introduction

The gold standard for the assessment of outcomes after hip and knee arthroplasty is prosthesis survivorship. However, modern advances in prosthetic design and technique are such that the threshold for joint arthroplasty has moved from salvage operations performed in extreme cases, to an intervention designed to improve the quality of life in patients who might otherwise cope without surgery. Hence, judging the success of the surgery may relate more to subtler improvements in quality of life, including relief of pain and improvement in function. Furthermore, technological innovation has improved the design of prostheses, ensuring survival in situ, barring infection, for at least a decade with relative certainty [19, 29, 47]. Consequently, the homogeneity of current prostheses (with respect to stable and lasting designs) has produced an emerging emphasis on quantifying subtler outcomes after arthroplasty.

Although there is some consensus as to which categories of outcome metrics should be applied to arthroplasty patients, there is no agreement as to which specific metrics are most appropriate. Instead, multitudes of metrics have been put forward in the literature and new metrics continue to be introduced. International researchers are subsequently forced to choose a metric based on its published psychometric properties, or, more alarmingly, based on precedence and extraneous political factors. This practice has led to significant variation in the reporting of outcomes after arthroplasty, particularly between nations. While general trends in outcomes can be contrasted with various outcome metrics, subtler differences in outcomes are lost in the psychometric variability between outcome tools. Ideally, international consensus should determine the most appropriate outcome metrics for wide scale employment. General approaches to standardizing outcome metrics for arthroplasty shall be discussed in this chapter.

Survival Analysis

Arthroplasty registries use prosthesis survival as the primary outcome. Survival analysis is a definitive metric that facilitates comparison of outcomes between nations. Currently, several national arthroplasty registries have the potential to compare and contrast survivorship outcomes [14, 19, 41, 46]. However, such comparisons are limited with respect to variation in demographics, including age at time of operation, diagnostic groupings, body mass index, gender, and activity levels. Research efforts should be directed at defining the demographics of each nation/center in detail so that the denominator of the comparative data can be determined. Without this level of research, comparison of outcomes in survivorship between nations/centers is prone to misinterpretation. Furthermore, the specific method of defining survivorship should be standardized [37]. For example, Cox's regression is a particularly useful method as it accounts for other factors such as age and gender, which are known to have an effect on outcomes. If such factors are not considered in outcome analyses, reported differences in survival curves between various prostheses are difficult to interpret, particularly on an international basis.

Arthroplasty registries function best as a surveillance tool for implant failure. As such, favorable and unfavorable trends in the outcomes of certain prostheses can be easily determined and disseminated back to the orthopedic community in a quality-improvement
feedback cycle. However, because arthroplasty registries are surveillance tools, there is an inherent lag in the reporting of outcomes, which creates a potential for sub-optimal implants or techniques to penetrate into and become part of the clinical norm prior to detection by the registry. A more accurate and predictive form of survivorship analysis would have the advantage of limiting new technology and techniques to fewer patients than is necessary to see trends with arthroplasty registries.

Radiostereometric Analysis

Radiostereometric Analysis (RSA) is a precise outcome tool that has accurate and reliable predictive ability with regard to implant aseptic loosening and, thus, survivorship [26, 48]. At the time of surgery, radio-dense tantalum markers are placed into the host bone and these beads are used to mark the implant. Post-operatively, biplanar simultaneous stereo X-rays of the joint are taken through a calibration cage with known fiducial points. The generated images are then imported into a RSA-software analysis package, and micromotion at the interface of the implant and host bone is calculated in three dimensions. These three vectors are combined into a metric of overall motion - maximum total point motion (MTPM). At six-month intervals, MTPM is plotted on the Y-axis against serial X-ray measures (on the X-axis) (Fig. 23.1). RSA curves follow a typical pattern: the implanted prosthesis either stabilizes with respect to MTPM over time, or it continues to migrate. If the prosthesis stabilizes, revision for aseptic loosening is unlikely. Conversely, if the prosthesis continues to migrate, revision for aseptic loosening is significantly more likely. The power of RSA is such that variation in MTPM patterns can be differentiated accurately as early as one year post-operatively, and as few as 30 to 40 patients need to be exposed to the experimental technology.

It is the author’s opinion that RSA is a critical technology for the rational development of minimally invasive surgery (MIS). As new MIS techniques and implants are introduced, earlier concerns over health-quality improvement and cost savings can be effectively addressed. National registries, while useful in overseeing the introduction of new technology, offer outcomes on a near real-time basis, and hence are not predictive. One theoretical concern regarding MIS is that the compromised exposure may lead to sub-optimal fixation and therefore reduced survivorship. RSA has the potential to predict the long-term survivorship of MIS implants, and, as such, allows for a more realistic determination of the economic model for MIS. What are the savings to the health-care system if hip-replacement patients are sent home on the same day of surgery, only to come to a premature revision? Questions like these can be most definitively answered with RSA.

Limitation of Survivorship Analysis

Arthroplasty registries rely on revision status as the sole endpoint for defining the outcome after arthroplasty. Revision status is a useful measure as it is relatively easy to define and the incidence of revision is definite. While definitive, revision status is a relatively blunt metric and is generally non-representative of the function, degree of pain relief, and overall patient satisfaction after knee arthroplasty. Furthermore, different surgeons have different thresholds for performing revisions and not all patients requiring revision surgery undergo the procedure because of co-existing medical problems, personal wishes etc. [12]. Revision status yields data only on the small minority of operations that fail [1]. The same set of arguments generally holds true for the outcome of continuous migration, as defined by RSA, which essentially acts as a surrogate for revision status. While there
is some evidence that subjective outcomes may be correlated with RSA-defined migration patterns, this phenomenon is not widely reported in the literature [20].

The Institute of Medicine defines health-care quality as “the degree to which health services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge” [38]. Previously, the “desired health outcome” of arthroplasty was for a prosthesis that performed in some minimal fashion to alleviate pain and improve function, as long as the prosthesis survived for a period of time without catastrophic complications. Currently, joint arthroplasty is a reproducible, effective and long-lasting procedure [19, 29, 47]. Consequently, comparative analyses of prosthetic models and surgical techniques are increasingly directed toward subjective arthroplasty outcomes.

**Subjective Outcomes**

**Background**

In 1947, the World Health Organization defined health as follows: “Health is not only the absence of infirmity and disease but also a state of physical, mental and social well-being.” Subsequently, in the surgical realm, the measurement of health moved from simply defining the success of a procedure by its effect on infirmity and disease, to defining what effect the intervention had on physical, mental and social well-being [12]. By this definition, it was no longer adequate to define the outcome of an arthroplasty, for example, by simply stating the prosthesis survival rate. This change in philosophy led to the development of general health-outcome questionnaires; examples that have been used in arthroplasty research include the 36-Item Short-Form Health Survey (SF-36) [4], the 12-Item Short-Form Health Survey (SF-12) [51], the Nottingham Health Profile (NHP) [23], and the Sickness Impact Profile (SIP) [40]. General health questionnaires focus on patients’ perceptions of their own health, including such diverse domains as ability to sleep, energy level, mood, and perception of body pain. These questionnaires are limited neither to specific disease nor patient cohort.

In an effort to avoid the surgeon bias associated with objective outcomes, additional disease-specific questionnaires were introduced to arthroplasty patients. Disease-specific questionnaires attempt to isolate the signal of interest by focusing questions around a particular disease state. In the 1980s, the Lequesne Index of Severity for the Knee (ISK) [33, 34] and the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) [2, 3] were introduced. Site-specific questionnaires attempt to isolate the signal in a similar fashion by focusing questions on a specific region of the body. The Oxford-12 Item Knee and Hip Scores (Oxford-12) were later developed and released in 1998 for use with knee and hip-arthroplasty patients as site-specific questionnaires [10, 11]. Patient-specific questionnaires, such as the Patient-Specific Index [52], use a novel approach to limit the noise within a questionnaire by asking patients to choose their own goals or objectives prior to an intervention and subsequently requesting that they rate or score how well those objectives have been accomplished. Global- or single-item questionnaires, such as those regarding patient satisfaction, are the most aggressive in their effort to limit noise by asking a single direct question regarding the state or condition of interest [44, 45]. Several authors have suggested that the simultaneous use of general-health- and disease/site-specific questionnaires yields complimentary data [17, 35, 39].

One aspect in the development and assessment of all subjective outcomes that is germane to health-care researchers is the concept of psychometrics. Psychometrics can be defined as “the scientific measurement of mental capacities and processes and of personality” [6]. In other words, psychometrics is the process that allows researchers to apply scientific methodology to the measurement of subjective outcomes. In practical terms, the published psychometric properties of a questionnaire pertain primarily to the validation of the questionnaire (i.e. how well the questionnaire measures what it is supposed to measure). The validation process usually involves three specific elements of questionnaire testing: validity, reliability, and responsiveness. Subjective outcome metrics that are to be applied in a valid fashion for health-outcomes research must meet these three basic criteria.

**Validity**

In contrast to validation, validity refers more specifically to how well the questionnaire measures the question of interest. The results are usually compared to a gold standard. Unfortunately, there is no gold standard for arthro-
Consequently, questionnaires for arthroplasty are usually validated against a postulated effect, or construct, that is an expected outcome of the intervention. Given that postulated constructs typically differ across nations, this practice presents a major limitation in the comparison of international subjective outcomes.

**Reliability**

Reliability refers to the ability of an outcome metric to remain unchanged when applied on two separate occasions in which no clinical change has occurred [12]. Reliability is the measure of the noise within a metric and can be described by the following equation:

\[
\text{Reliability} = \frac{\text{Subject Variability}}{\text{Subject Variability} + \text{Measurement Variability}}
\]

In order for an outcome metric to have acceptable reliability, it must have limited measurement variability.

**Responsiveness**

Responsiveness is a measure of a questionnaire’s ability to detect change when it is applied on separate occasions and a clinically significant change has occurred [12]. There are several methods of determining responsiveness, including the standardized effect size (SES). SES is calculated by subtracting the results of a questionnaire at time two from the results of the same questionnaire at time one and dividing the difference by the standard deviation of the test results from time one. Time one and time two represent a period over which a clinically significant change should have occurred, such as before and after arthroplasty surgery. A SES of 0.2 is considered small, 0.5 as moderate and greater than 0.8 as large [36].

Knee and hip arthroplasty have been shown to have a major impact on health-related quality of life when comparing pre-operative to post-operative status [10, 11, 32, 42, 43]. Dawson et al. have shown a SES of 2.0 for knee arthroplasty when the Oxford-12 Item Knee Score was applied pre- and post-operatively [11]. Such a SES can be considered profound, especially when a SES of 0.8 is considered large. Given the likelihood that such a large signal would mask subtler differences in outcome, an unusually large SES makes pre- and post-operative comparisons of different prosthetic designs or surgical techniques difficult to interpret and potentially irrelevant (Fig. 23.2). Paradoxically, the signal for pre- and post-operative comparisons after hip and knee arthroplasty is so loud (large) that it in effect functions as noise and obscures the subtler signal of interest [12]. Therefore, it may be more relevant to calculate responsiveness using an alternative method and/or to follow arthroplasty patients longitudinally between time two (a defined post-operative period) and time three. In this case, the large signal of the operative intervention would not obscure the subtler signal of interest. This is particularly relevant when comparing international outcomes as the signal will be further degraded by regional variation in demographics, techniques, implants, etc.

**Objective Outcomes**

A word about surgeon-defined objective outcome metrics is warranted. Initially, as surgeons began to realize a need to formally assess the results of their interventions, they began to derive outcome metrics that incorporated
various parameters, including technical procedure-specific outcomes (e.g. alignment, range of motion, etc.) as well as subjective patient-specific factors (e.g. pain). Examples of such assessment tools include the Hospital for Special Surgery Knee Score (HSS) [25], the Knee Society’s Clinical and Functional Scoring System (KSS) [24], and the Harris Hip Score (HHS). Unfortunately, and despite their continued popularity, the HSS and KSS scores have never been validated using formal psychometric validation procedures. Moreover, these questionnaires have been found to be exceedingly unreliable [49], leading some authors to conclude that these scoring systems should not be used [30]. The HHS has been validated for use, but it also relies on the objective and potentially biased input of the surgeon [50]. The psychometric robustness of these outcome metrics is limited, and caution should be used in their interpretation, particularly the summary scores that are produced. It is the author’s opinion that if these metrics are to be employed, summary scores should be avoided in analyses. Instead, individual sections within each metric should be reported and compared. Finally, given the lack of validated outcomes, these metrics should not be employed when comparing international outcomes.

Sources of Bias when Assessing Outcomes

The assessment of outcomes after hip and knee arthroplasty is subject to bias from several sources. Firstly, patient demographics may influence the results of questionnaire scores. Advanced age (>85 years) has been shown to have an adverse affect on subjective assessments after arthroplasty, as has low socioeconomic status, at least in North America [5, 7]. Gender has also been found to affect the results of health outcome questionnaires, particularly when used in association with hip or knee arthroplasty. Women tend to report greater pain and physical function limitation after hip or knee arthroplasty [27]. Co-morbidity has also been shown to adversely affect the results of knee arthroplasty, as assessed by questionnaire, for both joint-related and medical problems [5, 12, 18]. Charnley was aware of the potential biasing effect of co-morbidity, which was largely the impetus for the Charnley co-morbidity classification proposed for hip arthroplasty [8]. Gender, age, and co-morbidity should be considered when comparing outcomes after hip or knee arthroplasty, particularly international outcomes.

The mode of administration also significantly biases the results of health-outcome questionnaires. When a questionnaire is self-completed by the patient after surgery, as opposed to being administered by the investigator, the resulting questionnaire scores have been shown to be significantly lower (worse) [22, 44]. In addition, non-responders to a self-administered postal survey on quality of life tend to report worse quality of life than responders when followed-up with a telephone survey [21].

Finally, it is inadequate to simply translate a questionnaire into another language [15, 16]. Instead, the translated version needs to be tested for psychometric and cultural equivalence, in order to be deemed valid.

Selecting Appropriate Outcomes

When comparing international results on the introduction of minimally invasive hip and knee arthroplasty, little consensus exists as to which outcomes should be used in order to optimize the comparison. However, it is possible to make some recommendations based on the arguments made above.

As with conventional arthroplasty procedures, all MIS hip and knee arthroplasty procedures should be recorded prospectively into an arthroplasty registry, preferably at the national level. Concomitantly, demographic data, including gender, body habitus, and co-morbidity data, should be recorded. Such data is vital for meaningful international comparisons. Linkages to parallel hospital administrative patient databases could preclude the need to collect this data directly from the surgeon. A significant proportion of MIS hip- and knee-arthroplasty procedures should involve RSA technology in order to assess long-term survivorship at an early stage and limit the exposure of unproven techniques and implants. Without such RSA studies, the true global economics of MIS arthroplasty, which includes the cost of revision as a risk ratio compared to conventional arthroplasty, could not be determined until the survivorship period had been realized.

Surgeon-derived outcome metrics, such as the KSS and HSS knee scores, should generally be avoided as their lack of validation makes international comparisons suspect. The individual components of these scores, however, such as alignment and range of motion, should continue to be collected. Variations in these parameters are significant between nations [9]. Whenever possible, the most accurate form of reproducible
measurement should be employed. International standardization would be prudent.

Given that certain outcomes are perception-based (e.g., cosmetic appearance), it is important to collect subjective outcomes for MIS arthroplasty so as to better delineate the patient’s perception. There are several categories of subjective outcomes including general health, disease-specific, site-specific, patient-specific and single-item global questionnaires. Unfortunately, there is scant consensus with respect to which outcome measures are most appropriate. Each author advocates his outcome measure over others using, at best, statistical methodology that makes direct comparison of measures difficult to interpret. International consensus is required and should be sought. A general health questionnaire should be incorporated into the assessment process, as it would assist in standardizing comparison between nations. The SF-36 is well validated and has been translated for numerous countries. Site-specific questionnaires, such as the Oxford-12 seem to be better suited for assessing the subjective outcome of arthroplasty than disease-specific questionnaires, although the WOMAC is very popular in North America [13]. Patient-specific indexes may or may not be more appropriate for use in international comparative outcomes; more research needs to be done around this type of outcome metric. Single-item global questionnaires have been shown to be valid and informative for assessing knee arthroplasty outcomes at a national level, and the simplicity of the satisfaction question may transcend cultural differences [44, 45]. When analyzing data from subjective outcomes, demographic variables must be accounted for, especially co-morbidity, if meaningful international comparisons are to be made. Above all, the metrics employed must be validated.

Conclusions

There is currently no consensus as to the type of outcome data that should be collected when comparing international results of MIS arthroplasty. Prosthesis survivorship, as defined by arthroplasty registries and RSA, should be reported. If international comparisons are to be made, it is important to record demographic data as well-defined and accurately measured clinical parameters. Consensus needs to be reached regarding which subjective outcomes should be used, with consideration given to those questionnaires that are best-suited to arthroplasty. Ongoing research is required.

References