From the Department of Orthopaedics
Lund University Hospital, SE-221 85 Lund, Sweden

Subjective outcomes after knee arthroplasty

Michael J. Dunbar

THESIS

ACTA ORTHOPAEDICA SCANDINAVICA SUPPLEMENTUM NO. 301, VOL. 72, 2001

List of Papers

This thesis is based on the following papers:

1. **Robertsson O, Dunbar MJ, Pehrsson T, Knutson K, Lidgren L.** Patient satisfaction after knee arthroplasty. A report on 27,372 knees operated on between 1981 and 1995 in Sweden. *Acta Orthop Scand* 2000; 71(3): 262-7

2. **Dunbar MJ, Robertsson O, Ryd L, Lidgren L.** Appropriate questionnaires for knee arthroplasty: Results of a survey to 3600 patients from the Swedish Knee Arthroplasty Registry. Accepted *JBJS (Br)* 2000

3. **Robertsson O, Dunbar MJ.** Patient satisfaction compared with general health and disease specific questionnaires in 3600 patients operated on with knee arthroplasty. Accepted *J Arthroplasty* 2000.

4. **Dunbar MJ, Robertsson O, Ryd L.** What's all that noise? The effect of co-morbidity on health outcome questionnaire results after knee arthroplasty. Submitted *J Arthroplasty* 2000.

5. **Dunbar MJ, Robertsson O, Ryd L, Lidgren L.** Translation and validation of the Oxford-12 Item Knee Score for use in Sweden. *Acta Orthop Scand* 2000; 71(3): 268-274

6. **Dunbar MJ, Valdivia GG, Parker DA, Ryd L, Bourne R, Rorabeck C.** Post-operative patient disposition after knee arthroplasty based on pre-operative WOMAC scores. *In Manuscript* 2000.

Definitions and abbreviations

Ceiling effect – The property of scoring the worst possible score on a questionnaire such that a repeated application would not be capable of demonstrating a worse score if the patient clinically deteriorated.

Domain – A sub-score within a questionnaire meant to cover a specific condition of interest, e.g., Body Pain, which is a domain within the SF-36.

Feasibility – The average usable response rate for a questionnaire when self-administered in a postal survey.

Floor effect – The property of scoring the best possible score on a questionnaire such that a repeated application would not be capable of demonstrating an improvement in score if the patient clinically improved.

ICC – Intraclass correlation coefficient, often used when assessing test-retest reliability on ordinal scales.

Imputation – Computer assisted completion of missing items from a questionnaire based on how associated items within the questionnaire were answered.

Item – A single question within a domain or questionnaire.

Lequesne – Lequesne Algofunctional Knee Index (site specific questionnaire).

Likert Scale – A rating scale in which raters express their opinions on a given subject by marking a box within a continuum of disagree-agree statements.

NCR – National Census Registry

NHP – Nottingham Health Profile (general health questionnaire).

Noise – Any part of an observation that does not contribute to the signal of interest. Often defined statistically as the variation between individuals within a study group, although sometimes the variation between individuals is the signal of interest in health outcomes research.

Outcome – The result or effect of a defined intervention.

Oxford-12 – Oxford-12 Item Knee Score (site specific questionnaire).

Patient burden – The amount of time and assistance required by a patient in order to complete a given questionnaire.

PIN – Personal Identification Number

Questionnaire (Disease Specific) – A questionnaire designed to measure an outcome in a patient population with a similar disease state.

Questionnaire (General Health) – A questionnaire designed to measure an outcome in a general patient population regardless of disease state.

Questionnaire (Site Specific) – A questionnaire designed to measure an outcome in a patient population regarding a specific joint involved in a disease process.

Reliability (internal consistency) – The extent to which items within a domain measure the same subject of interest.

Reliability (test-retest) – The property of a questionnaire that yields the same or similar score when applied on repeated applications and no clinically relevant change has occurred.

Response rate – The percentage of questionnaires returned by patients who were assumed to be alive and living at the address to which the questionnaire was sent.

Responsiveness – The property of a questionnaire that yields different scores when applied on repeated applications and a clinically relevant

change has occurred.

Revision – The addition, exchange, or removal of an endoprosthetic knee component

ROC curve – Receiver Operating Characteristic Curve

SF-12 – 12-Item Short-Form Health Survey (general health questionnaire).

SF-36 – 36-Item Short-Form Health Survey (general health questionnaire).

Signal – The part of an observation that forms the relevant part of any measurement (as opposed to noise)

SIP – Sickness Impact Profile (general health questionnaire).

SKAR – The Swedish Knee Arthroplasty Registry

Skew – The extent to which a frequency distribution deviates from a normal distribution.

TKA – Total knee arthroplasty.

UKA – Unicompartmental knee arthroplasty.

Validity – The extent to which a questionnaire appropriately measures the condition of interest.

Validity (construct) – The extent to which a questionnaire correlates to a theoretical model (construct) that also measures the condition of interest.

Validity (content) – The extent to which a questionnaire covers the condition of interest.

Validity (criterion) – The extent to which a questionnaire correlates to the "gold standard" (criterion) that also measures the condition of interest.

WHO – World Health Organization.

WOMAC – Western Ontario and MacMaster Universities Osteoarthritis Index (disease specific questionnaire).

Introduction

Historical background

Knee arthroplasty as related to outcomes

The first published report on endoprosthetic knee arthroplasty is often attributed to Gluck (1890). Gluck employed endoprostheses made of ivory for the treatment of knee joints destroyed by tuberculosis. At the time, the only alternatives to this "radical" intervention were amputation, arthrodesis, interpositional arthroplasty, or benign neglect. Faced with such severe joint disorders, Gluck's surgical interventions were initially deemed successful, mostly because the alternatives to the prosthesis were so dismal. Still, Gluck later cautioned about the use of this prosthesis because of continued problems with infection. This note of caution represented the first report on the outcomes after endoprosthetic knee arthroplasty.

Perhaps because of the warnings from Gluck, interpositional arthroplasty continued as a standard of treatment for severely diseased knee joints. Interpositional materials included pigs' bladders, fascia lata, patellar bursae, vitallium covers, and cellophane (Shiers 1954). In 1949, Speed reported on the outcome of 65 interpositional arthroplasties and graded them as good (n = 29), fair (17), poor (6) and failures (13) (Speed et al. 1949). Miller reported on 37 interpositional arthroplasties in 1952, which demonstrated worse results than Speed (Miller et al. 1952). 11 were reported as good, 8 as fair and 18 as failures. These outcome metrics were surgeon derived and did not rely on input from the patients.

In the face of such poor results and with the continued development of modern anesthesia, aseptic technique and antibiotic prophylaxis, the modern era of endoprosthetic knee arthroplasty began. Shiers reported a case study of 2 patients using a stainless steel hinged prosthesis (Shiers 1954). In 1 patient, heterotopic ossification limited the results, but the other was deemed to be successful. Shiers considered the operation a success because the patient was painless, could walk without a stick, and could ascend and descend stairs. Walldius reported encouraging results of endoprosthetic knee arthroplasty using a cobalt-chromium hinged prosthesis (Walldius 1957, reprinted 1996). Although no formal scoring systems were applied in these studies, the authors did consider subjective and objective outcomes in the determination of the success of the operation.

Gunston, the originator of an endoprosthesis consisting of individual stainless steel semicircular runners articulating with separate high density polyethylene runners cemented to the tibia (The Polycentric Knee), reported on the results of 22 knee arthroplasties in 20 patients (Gunston 1971). With 2 years follow-up, Gunston reported on the radiographic results as well as pre and post-operative pain, flexion, and lateral instability. Whether or not the mobility of the patient had improved or was unchanged as well as a report of complications was recorded. This assessment began to resemble some of the current outcome tools used to assess knee arthroplasty. Interestingly, Gunston did not summarize the variables nor produce a score, but instead reported each parameter on its own merits.

In the early 1970's Swanson and Freeman designed an unlinked duocondylar prosthesis with a metal-on-polyethylene articulation which was cemented to the bone (Freeman et al. 1986). In 1972, the prosthesis was modified to include a patellar component that articulated with the femoral component as well as a stemmed tibial component. This prosthesis was referred to as the Total Condylar Knee (Insall et al. 1979). At approximately the same time, springing from the work of Gunston, less constrained unicompartmental prostheses were introduced. These included the Marmor and St. Georg Sledge (Engelbrecht 1971, Marmor 1973). The introduction of these prostheses resulted in relatively predictable outcome after knee arthroplasty. Current knee prostheses can directly derive their lineage from these prostheses and represent variations of the basic concepts introduced.

The importance of the advances in prosthetic design relates directly to the fact that the threshold for endoprosthetic knee arthroplasty had moved from that of a salvage operation performed in extreme cases, to an intervention designed to improve the quality of life in patients who might otherwise cope without the intervention. Hence, judging the success of the intervention may relate more to subtler improvements in quality of life, including relief of pain and improvement in function. Furthermore, current prostheses have all benefited from the technological learning curve in the design of prostheses, and modern prostheses can be expected to survive in situ, barring infection, for at least a decade, or perhaps 2 decades, with relative certainty. The net effect of the homogeneity of current prostheses (with respect to stable and lasting designs) has been for an emerging emphasis on somehow quantifying subtler outcomes after knee arthroplasty.

Objective outcomes

With the advent of prosthetic components that demonstrated predictably good results, it became evident that more formalized outcome metrics were necessary. The initial response was for surgeons to assess the results of their interventions. In 1976, Insall et al. introduced a surgeon derived outcome score for knee arthroplasty that incorporated various parameters including technical outcomes related to the procedure (e.g. alignment, range of motion, etc.) and subjective patient factors such as pain (Insall et al. 1976). This questionnaire has come to be known as the Hospital for Special Surgery Knee Score (HSS). In 1989, Insall et al. developed a second surgeon derived score, which incorporated similar parameters. This score has come to be known as the Knee Society's Clinical and Functional Scoring System (KSS) (Insall et al. 1989). The HSS and KSS have been used fairly extensively in outcome studies on knee arthroplasty (Amendola et al. 1989, Joseph et al. 1990, Armstrong et al. 1991, Nafei et al. 1993, Fehring et al. 1994, Hirsch et al. 1994, Knight et al. 1997, Barrack et al. 1998). Unfortunately, and despite their continued popularity, the HSS and KSS scores have never been validated using formal psychometric validation procedures. Furthermore, these questionnaires have been found to be exceedingly unreliable (Ryd et al. 1997), leading some authors to conclude that these scoring systems should not be used (Konig et al. 1997).

Subjective outcomes

Pythagoras mused that "man is a measure of all things" (Strohmeier et al. 1999). The implication of this statement speaks to the conceptualization that the distinction between mind and body is blurred, or indeed that there is no distinction at all. While the Western philosophical distinction between mind and body has its origins from the ancient Greeks, it was the works of Renés Descartes that formalized the modern distinction between mind and body (Descartes 1986). According to Descartes, the rational soul is an entity distinct from the body that may or may not be aware of the signals passing through the body via the interfibrillar spaces. The interfibrillar spaces (i.e. sensory nervous system) were "extended" into the physical world, while the rational soul (i.e. consciousness) was not. This distinction between mind and body has persisted into modern Western medical thought.

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." This definition reintroduced the concept that the mind and body are in fact one, and the "well being" of the mind and body combined represents health. Subsequently, the measurement of health moved from simply defining the success of a procedure by defining its effect on infirmity and disease, to the more ambitious approach of defining what effect the intervention had on physical, mental and social well being. By this definition, it was no longer adequate to define the outcome of a knee arthroplasty, for example, by simply stating what the range of motion was or what the impact was on mobility, such as Gunston and other innovators had done, as mentioned above. Instead, a more comprehensive metric was needed.

The definition of health by the WHO was perhaps the impetus for the modern movement to measure physical, mental and social well being. The first attempts at quantifying general health were with single-item global ratings which were

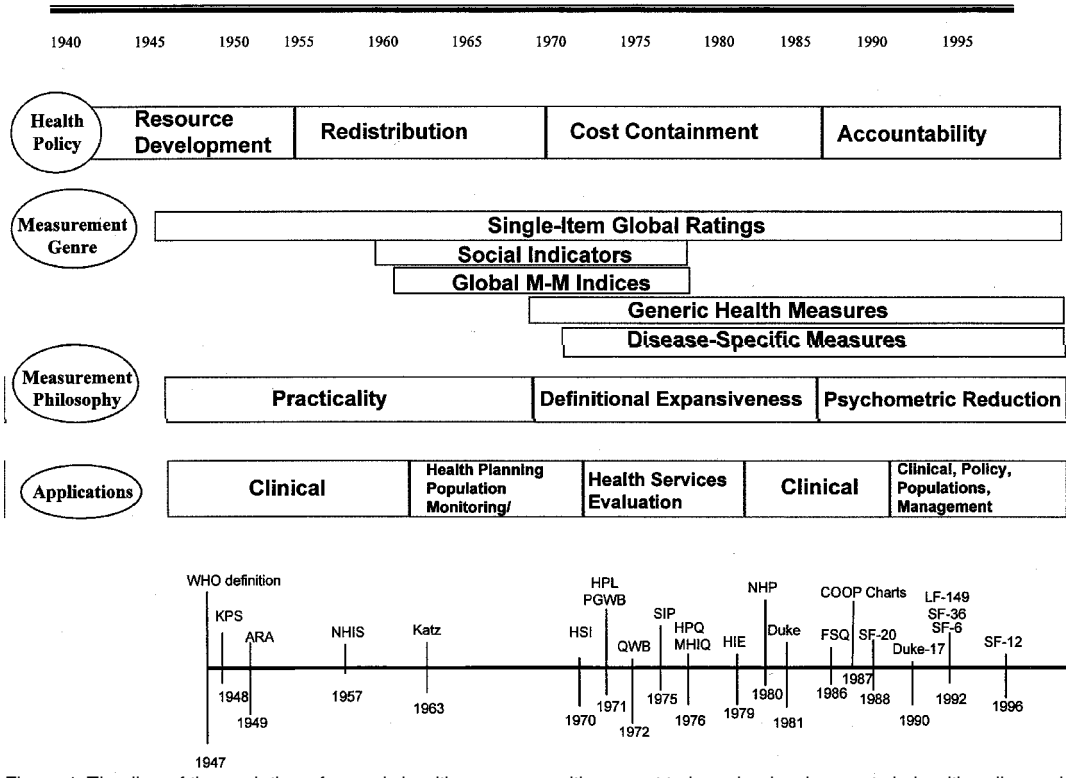

Figure 1. Timeline of the evolution of generic health measures with respect to broader developments in health policy and health status assessment. ARA = American Rheumatoid Association Functional Class; COOP = Dartmouth COOP Poster Charts; Duke = DukeUNC Health Profile; Duke-17 = Duke Health Profile; FSQ Functional Status Questionnaire; HIE = Health Insurance Experiment; HPL = Human Population Laboratory; HPQ = Health Perceptions Questionnaire; HS1 Health Status Index; KPS = Karnofsky Performance Status; Katz = Katz Index of Activities of Daily Living; LF-149 = Medical Outcomes Study 149-Itern Functioning and Well-Being Profile; M-M = morbidity and mortality; MHIQ = McMaster Health Index Questionnaire; NHIS = National Health Interview Survey; NHP = Nottingham Health Profile; PGWB = Psychological General Well-Being Scale; QWB = Quality of Well-Being Scale; SF-6 = Medical Outcomes Study 6-Item Health Survey; SF-12 = Medical Outcomes Study 12-Item Health Survey; SF-20 = Medical Outcomes Study 20-Item Health Survey; SF-36 = Medical Outcomes Study 36-Item Health Survey; SIP = Sickness Impact Profile; WHO = World Health Organization. Reprinted with permission from Annals of Internal Medicine (McHorney 1997).

designed to augment organ specific or more physiological outcomes. With time, a large number of questionnaires were developed that asked more questions around various aspects of health, such that separate scores for each of these health domains were generated. Domains that attempted to account for physical, mental and social well being included Emotional Reaction, Sleep, Social Isolation, Body Pain, and Social Functioning, for example. Advanced study and refinement of these tools continues today. The introduction and evolution of generic (or general) health measurements has been well documented by McHorney (1997), and can be represented graphically (Figure 1). Measurements of this sort are often referred to as "subjective" and are difficult to quantify. Still, some form of logical metric was imperative for further research. This dilemma was eloquently alluded to by Lord Kelvin when he said, "I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind." (Thompson 1910). The WHO continues to be interested in this area of outcomes research. At a recent workshop in January 2000 under the umbrella of the Bone and Joint Decade 2000–2010 the need to standardize outcome metrics for musculoskeletal research was discussed (http://www.bonejointdecade.org/).

While the WHO definition of health may be

largely responsible for the emergence of general health outcome questionnaires, the first aspect of the definition, i.e. "…the absence of infirmary or disease…" has not been lost on researchers. A similar evolution in health outcome questionnaires focused on the organ (or site) or physiologic process (disease) has come about. This work has its roots in the very early reports of Gluck and Gunston, who made some effort to quantitate the outcomes of their specific intervention, at the joint and/or disease level, as mentioned above. This was followed with the biased surgeon-derived HSS and KSS, also mentioned above.

Partly in an effort to avoid the surgeon bias associated with objective outcomes, other disease/site specific questionnaires emerged that were relevant to knee arthroplasty. In the 1980's the Lequesne Index of Severity for the Knee (ISK) (Lequesne et al. 1987, Lequesne 1989) and the Western Ontario and MacMaster Universities Osteoarthritis Index (WOMAC) (Bellamy et al. 1984, Bellamy et al. 1988) were introduced. The Oxford-12 Item Knee Score (Oxford-12) was later developed and released in 1998 to be used specifically with knee arthroplasty patients (Dawson et al. 1998). Unlike the HSS and KSS, these questionnaires do not rely on surgeon input and all have been well validated.

The Swedish Knee Arthroplasty Registry

The Swedish Knee Arthroplasty Study was initiated in 1975 by the Swedish Orthopaedic Society (Robertsson et al. 1999c). The result of this initiative was the Swedish Knee Arthroplasty Registry (SKAR) which has prospectively registered knee arthroplasties since 1975 and currently has data on over 70,000 knee operations (http://www.ort.lu.se/knee/). The SKAR represents the first national health care quality register ever. In Sweden alone there are now over 100 national registries which record data on all kinds of health interventions. Initially, endoprosthetic knee arthroplasty was a relatively uncommon procedure and an ambitious effort was made to collect radiographic data, a surgeon completed questionnaire and a modified translation of the British Orthopaedic Association Knee Assessment Chart (Aichroth et al. 1978). This schedule for data collection soon proved unwieldy as the incidence of knee arthroplasty rapidly increased. Furthermore, the comprehensiveness of the data collection came at the expense of voluntary contribution to the SKAR. Subsequently, a decision was made to scale back the data collected to key demographic and implant related factors, as well as to use revision as the single definitive endpoint. Outcome questionnaires were no longer part of the data collected with the SKAR.

In 1982, Tew et al. described a method of survival analysis for knee arthroplasty which made it possible to estimate the annual failure rate and the cumulative 10 year survival rate (Tew et al. 1982). Since 1985, the SKAR has used survivorship methods for evaluating outcomes after knee arthroplasty, with revision as the endpoint. Initially, life table curves were generated using the Wilcoxon, log-rank and other similar tests. Cox's regression was later used by the SKAR because of the inability of the above mentioned tests to account for other factors, such as age and gender, that are known to have an effect on outcomes. Without accounting for such factors, reported differences in survival curves between various prostheses were difficult to interpret (Robertsson 2000).

Today, the SKAR is somewhat unique because of its completeness and length of follow-up. In essence, the database represents a nation's experience with knee arthroplasty since its modern inception. The effect of the longevity and completeness of follow-up, facilitated with the use of a national personal number, has afforded effectual observations regarding various aspects of knee arthroplasty (Knutson et al. 1984, Knutson et al. 1985, Bengtson et al. 1986, Bengtson et al. 1989, Bengtson et al. 1991, Lewold et al. 1993, Lewold et al. 1996, Robertsson et al. 1997, Lewold et al. 1998, Robertsson et al. 1999d). The SKAR has also formed the basis for a number of PhD dissertations (http://www.ort.lu.se/knee/engversion/disertationseng.html).

The SKAR has relied on revision status as the sole endpoint for defining the outcome after knee arthroplasty. This has particular merits as an outcome metric as it is relatively easy to define and the incidence of revision is definite. The SKAR has defined revision as the addition, removal, or exchange of an endoprosthetic component, including amputation (Robertsson et al. 1999c).

Revision status within the SKAR has been demonstrated to be accurate (Robertsson et al. 1999b). While definitive, revision status is a relatively blunt metric and is generally non-representative of the functional performance, 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. Revision status yields data on the small minority of operations that fail and tells us nothing of the status of the majority of patients who have not come to revision (Apley 1990). Finally, revision status does not speak directly to the "…physical, mental and social well being of the patient", as outlined in the WHO definition of health. Indeed, revision status does not even directly address the "…absence of infirmary or disease…" aspect of the definition, as it is not clear as to what impact revision has on these aspects of the definition.

Impetus for assessing outcomes utilizing the Swedish Knee Arthroplasty Registry

The Institute of Medicine defines health care quality as "the degree to which health services for individuals and populations increases the likelihood of desired health outcomes and are consistent with current professional knowledge" (Palmer 1997). In the time of Gluck and even Gunston, the "desired health outcome" of knee arthroplasty was for a prosthesis that performed in some minimal way to alleviate pain and improve function, as long as the prosthesis survived some minimal time without catastrophic complications. Currently, endoprosthetic knee arthroplasty is a reproducible, effective and long lasting procedure (Knutson et al. 1986, Knutson et al. 1994, Robertsson et al. 1999d). Subsequently, when comparing various prosthetic models, surgical techniques, etc. for knee arthroplasty, the degree to which knee arthroplasty increases the likelihood of desired health outcomes relates more to subjective and qualitative outcomes. This is the impetus for the application of subjective health outcome questionnaires to the SKAR.

Subjective health outcome questionnaires

Psychometric considerations

Psychometrics can be defined as "the scientific measurement of mental capacities and processes and of personality " (Brown 1993). 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 mostly to the validation of the questionnaire, or, defining how well the questionnaire measures what it is supposed to measure, in a global sense. The validation process usually involves three specific aspects of questionnaire testing: validity, reliability, and responsiveness.

Validity refers more specifically (as opposed to validation) to how well the questionnaire measures the question of interest. Validity can take many forms and numerous synonyms have been utilized in conjunction with it. Theses include criterion, construct, convergent, divergent, and content validity. In order to comment on the validity of a questionnaire, the results of the questionnaire must be compared to something.

Criterion validity refers to the comparison of the metric to a "gold standard". For example, a thermometer is the gold standard for measuring body temperature. If a questionnaire was designed to measure body temperature, the items within may inquire about how warm the patient felt, whether or not they had chills, etc. The results of this questionnaire could be directly correlated to the gold standard (criterion). Unfortunately, there is no gold standard for knee arthroplasty (Kirshner et al. 1985, Kreibich et al. 1996). Consequently, questionnaires for knee arthroplasty are usually validated against a postulated effect that should result from the intervention. Such a postulation is referred to as a construct.

Construct validity may be determined against another previously validated questionnaire or a consensus statement, for example. Divergent and convergent validity can be used as a check for the construct in that items within a questionnaire that relate to knee function, for example, should improve after knee arthroplasty (convergent), while items that are not related to the knee, such as eating, should not change (divergent).

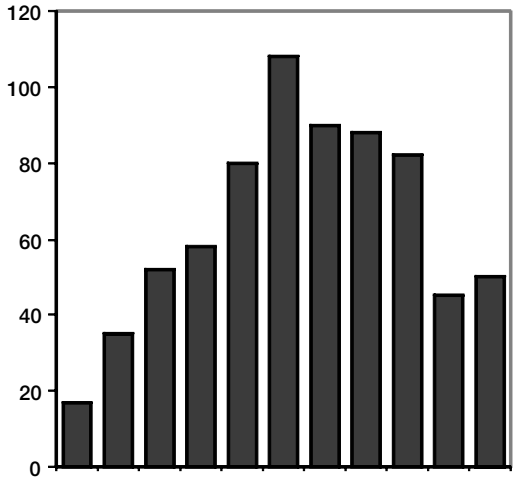

Figure 2a. Frequency distribution of scores for the Vitality domain of the SF-36 demonstrating a near Normal distribution with relatively few patients reporting the lowest possible (floor effect) or the highest possible (ceiling effect) scores.

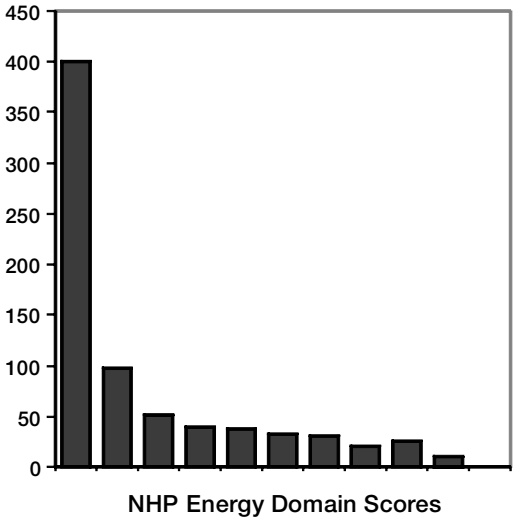

Figure 2b. Frequency distribution of scores for the Energy domain of the NHP (comparable to the Vitality domain of the SF-36) demonstrating a skewed distribution with the majority of patients reporting the lowest possible score (floor effect).

A note of caution is warranted when considering construct validity. Construct validity in the absence of a gold standard, such as the case with knee arthroplasty, is problematic. Often, questionnaires are validated against another questionnaire that has previously been validated. Further investigation may reveal that the previously validated questionnaire has been validated against a construct. Hence, a circuitous logical argument can be associated with outcome questionnaires with potential sophistic implications. There is no "*cogito ergo sum*" on which to base construct validity in the absence of a gold standard.

Content validity addresses whether a questionnaire has enough items and adequately covers the domain of interest (Streiner et al. 1998). For example, if a questionnaire is designed to measure how much mobility a patient has gained from a knee arthroplasty intervention, then by inference, a patient that scores well on the questionnaire could be assumed to have good mobility. However, if the items within the questionnaire do not ask specifically about mobility, then the inference is invalid (not necessarily the questionnaire). Questionnaires with good content validity cover the target behavior well and subsequently provide for valid inferences. Content validity can be tested by investigating the frequency distribution of the scores produced by a questionnaire or the domains within. In particular, the floor and ceiling effect are important when assessing content validity. A floor effect occurs when a respondent scores the lowest (i.e. best) possible score on a questionnaire. Thus, if a patient were to clinically become better, the questionnaire would be unable to reflect that change. The content of the behaviour would not be covered and inferences would be invalid. The same argument holds true for ceiling effect, which occurs in an opposite direction (Figures 2a and 2b).

Reliability refers to the ability of an outcome metric to remain unchanged when applied on two separate occasions and no clinical change has occurred. Essentially, in its most basic sense, reliability is the measure of the noise within a metric and can be conceptualized by the following equation:

$$\text{Reliability} = \text{Subject variability} / (\text{Subject variability} + \text{Measurement variability})$$

In order for an outcome metric to have acceptable reliability, it must, by the definition proposed here, have limited measurement variability.

Outcome metrics have been criticized because of the perception that they yield "soft" data, at

least in comparison to more standardized technological laboratory tests that permeate the medical field, such as serum potassium, or hemoglobin. Such tests are felt to yield "hard" data as the methodology for such tests is well described, the precisions are high and the reproducibility is excellent. Still, the perception that questionnaires yield only soft data must not prevent the clinically relevant questionnaire data from being utilized as this data, perhaps more so than any other, speaks to the humanistic side, or art, of medicine. Such an argument has been well described by Feinstein when he said the following: "If we say that cardiac size became smaller, that cardiac rhythm became normal, and that certain enzyme levels became normal, the description could pertain to a rat, a dog, or a person. But if we say that chest pain disappeared, that the patient was able to return to work, and the family was pleased, we have given a human account of human feelings and observations."

Classically, the test-retest reliability of an outcome metric is investigated by determining the Intraclass Correlation Coefficient (ICC) (Bland et al. 1996). The ICC is advantageous over other correlation coefficients, such as Spearman or Pearson, as it is not biased by the order in which pairs of data are compared. Subsequently, learning effects that may occur when a questionnaire is applied on two separate occasions will not influence the ICC. An ICC value between 0.60 and 0.79 can be considered as fair, 0.80 to 0.89 as good and 0.90 and above as excellent. Test-retest reliability values greater than 0.90 are required if consideration is being given to employing a questionnaire in a discriminative application on a patient-to-patient basis, as opposed to discriminating between groups (Ware et al. 1992).

Test-retest reliability is related to the number of items within a questionnaire as the true variance will increase as the square of the number of items, while the error variance will increase linearly with the number of items (Streiner et al. 1998). Generally then, the greater the number of items within a questionnaire, the better the test-retest value will be. This may have implications for questionnaire selection when good test-retest reliability is required, given the large variation in the number of items per questionnaire. Item reduction comes at the expense of test-retest reliability.

Reliability can also be investigated using Cronbach's Alpha statistic (Cronbach 1955, Bland et al. 1997). Cronbach's Alpha addresses the homogeneity of the items (questions) within an outcome questionnaire domain or total score and is complimentary to the ICC as a metric of reliability. Cronbach's Alpha is used primarily in the development of a questionnaire as a means of reducing the number of items within a scale as the statistic determines the inter-item correlation for each item within a domain. A value from 0 to 1 is produced with a value of 0.60 to 0.79 indicative of fair internal consistency, 0.80 to 0.89 as good internal consistency, and greater than or equal to 0.90 as excellent internal consistency (Feinstein 1987). Cronbach's Alpha is calculated n times for a scale (n = number of items within the scale) with 1 item omitted each time. If the value for Cronbach's Alpha increases with the omission of an item, then that item can be argued to be deviating from the area of interest inquired about within the scale and can therefore be omitted from the finalized scale. Cronbach's Alpha is used when the items within a scale are polychotomous. Dichotomous items, such as in the NHP, require a variation of Cronbach's Alpha known as the Kuder Richardson Formula 20.

As alluded to above, health outcome questionnaires have been criticized for yielding soft data and the softness or hardness of data is generally referring to the reliability of the questionnaires (both the ICC and Cronbach's Alpha). However, when evaluating relevant health outcome questionnaires on a target population, questionnaires have been shown to demonstrate fair to excellent reliability and therefore can be considered relatively hard. Generally, disease/site specific questionnaires produce harder data than general health questionnaires (Figures 3a and 3b). Some "hard" and "objective" data yield distinctly poor ICC values, making them actually rather "soft" (Ryd et al. 1997).

Responsiveness is a measure of a questionnaires ability to detect change when it is applied on separate occasions and a clinically significant change has occurred between applications. By definition, responsiveness is related to a longitudinal application of a questionnaire, however, as

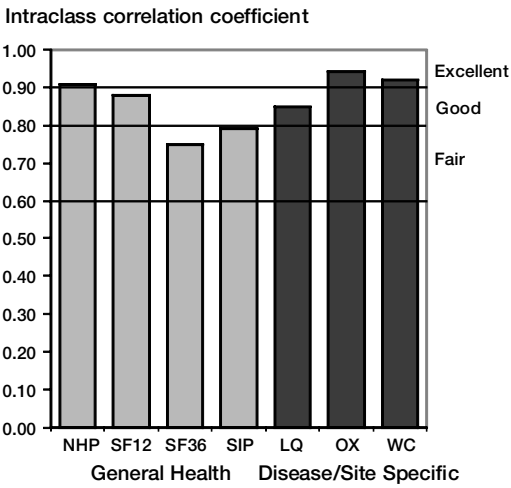

Figure 3a. Intraclass correlation coefficient values for test-retest reliability results of four general health and three disease/site specific questionnaires. All questionnaires tested demonstrate at least "Fair" test-retest reliability.

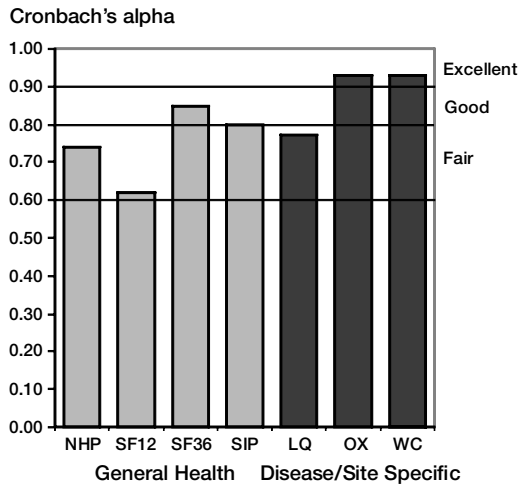

Figure 3b. Cronbach's alpha values for internal consistency reliability of four general health and three disease/site specific questionnaires. All questionnaires tested demonstrate at least "Fair" internal consistency reliability.

outlined above, the purpose of this study was to define appropriate questionnaires for cross-sectional discriminative application. Nevertheless, determining a questionnaire's responsiveness is integral to the validation process. Although responsiveness may have been previously defined for a questionnaire, often the investigations have been performed on dissimilar populations; therefore investigating responsiveness on the target population is necessary. Questionnaire validation is a dynamic unending process (Nunnally et al. 1994).

There are several methods of determining responsiveness, including the standardized effect size (Deyo et al. 1986, Guyatt et al. 1987, Kreibich et al. 1996, Essink-Bot et al. 1997, Wright et al. 1997). Standardized effect size is calculated by subtracting the results of a questionnaire at time 2 from the results of the same questionnaire at time 1 and dividing the difference by the standard deviation of the test results from time 1. Time 1 and time 2 represent a period over which a clinically significant change should have occurred, such as before and after a therapeutic intervention, be it a drug therapy or surgery, for example. A standardized effect size of 0.2 is considered small, 0.5 as moderate and greater than 0.8 as large (Meenan et al. 1991).

Knee and hip arthroplasty have been shown to have a major impact on health related quality of life when comparing preoperative to postoperative status (Laupacis et al. 1993, Rissanen et al. 1995, Ritter et al. 1995, Dawson et al. 1996b, Dawson et al. 1998). In fact, Dawson et al. have shown a standardized effect size of 2.0 for knee arthroplasty when the Oxford-12 Item Knee Score was applied pre- and postoperatively (Dawson et al. 1998). Such a standardized effect size can be considered profound, especially when a standardized effect size of 0.8 is considered large. Such profound results make pre- and postoperative comparisons of different prosthetic designs, surgical techniques, etc. using a given questionnaire difficult to interpret and potentially irrelevant as the assumed subtle differences in questionnaire results would be lost in the large signal. Paradoxically, the signal for pre- and postoperative comparisons after knee arthroplasty is so loud (large) that it in effect functions as noise and obscures the subtler signal of interest. Therefore, it may be more relevant to calculate responsiveness using an alternative method and/or to follow arthroplasty patients longitudinally between time 2 (a defined postoperative period) and time 3. In this case, the large signal of the operative intervention would not obscure the subtler signal of interest.

The Receiver Operating Characteristic Curve (ROC Curve) has been shown to be of value as a surrogate to classic responsiveness measures

when longitudinal data is not available (Hanley et al. 1982, Deyo et al. 1986, Centor 1991, Essink-Bot et al. 1997). This is particularly relevant for the reasons listed above and because the SKAR to date has not applied questionnaires in a longitudinal fashion. The ROC Curve method has its origins from the operation of radar equipment during the Second World War. At that time, the radar operators, and others, were interested in optimizing the signal to noise ratio of their receivers. Initially, as the gain on the equipment was increased, the signal correspondingly increased rapidly. However, at some point, the gain in the noise was greater than the gain in the signal. This represents the "cut-point" of interest and essentially the cut-point represents the dichotomization of continuous data. To construct a ROC Curve the true positive rate (sensitivity) of a test is plotted on the Y-axis and the false positive rate (1-specificity) is plotted on the X-axis. These two values are determined for each possible cut-point and a curve is subsequently generated. The area under the ROC Curve is used as a gauge of the discriminative ability of the test, with an area of 1.0 representative of a perfectly discriminative test and an area of 0.5 as a non-discriminative test. An example of a ROC Curve is demonstrated in Figure 4. In this case, Questionnaire A has better discriminative ability than Questionnaire B.

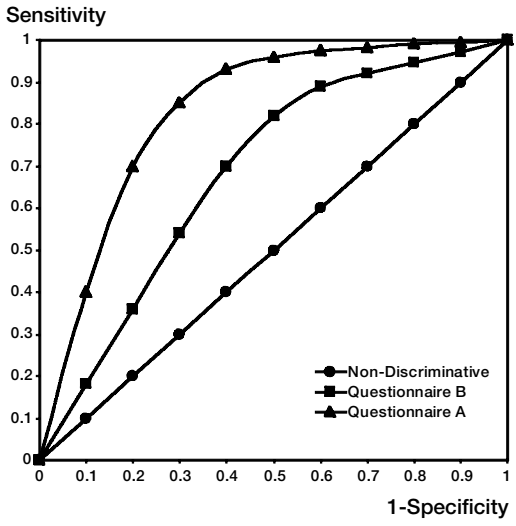

Figure 4. Example of two possible Receiver Operating Characteristic Curves (Questionnaire A and B). The area under the curve is directly related to the discriminative ability of the questionnaire. In this example, Questionnaire A has better discriminative ability than Questionnaire B.

Specific limitations related to the Swedish Knee Arthroplasty Registry

The large number of patients registered with the SKAR makes it impractical for a comprehensive questionnaire application to be performed in any format other than a postal survey. Subsequently, any questionnaires used would have to be completed solely by the patient without input from a health care provider. Ethically, imposing such questionnaires on patients should result in minimal patient burden. Patient burden, for the purposes of this study, refers to the time required for a patient to complete any given questionnaire and the requirement for patients to seek help in completing the questionnaires. Associated with patient burden is feasibility of the postal survey. Feasibility refers to the percentage of questionnaires returned multiplied by the number of those questionnaires that were returned completed. It could be hypothesized that simple, shorter questionnaires would impose fewer burdens and would therefore have higher feasibility than longer, more elaborate questionnaires. This hypothesis has not been definitively investigated in the literature. This is compounded by the fact that patients registered with the SKAR tend to be elderly. Burden and feasibility therefore is more of an issue with this unique population than an average general population sample.

Another limitation associated with the SKAR relates to the fact that preoperative health outcome questionnaires are not available for comparative purposes. Therefore, any questionnaire applied would have to function in a discriminative fashion. Technical differences in the development and construction of questionnaires may make them more or less favourable for a discriminative application (Kirshner et al. 1985). Most questionnaires have not been validated while accounting for this.

Questionnaires used with the SKAR need to be available in a translated and validated Swedish language version. It is inadequate to simply translate a questionnaire into another language (Guillemin et al. 1993, Guyatt 1993). Instead, the translated version needs to be tested for psychometric and cultural equivalence, in order to be deemed valid.

Finally, the limitations listed here are relevant for other large national databases in Sweden and elsewhere. The exception is, of course, the need for a Swedish version of the questionnaire. Instead, the questionnaire needs to be available in a native language form.

Sources of bias when assessing outcomes

Health outcome questionnaires are subject to bias from several sources. Firstly, patient demographics may influence the results of questionnaire scores. Advanced age (greater than 85 years) has been shown to have an adverse affect on subjective assessments after knee arthroplasty, as has low socioeconomic status, at least in North America (Callahan et al. 1994, Brinker et al. 1997). Gender has also been found to affect the results of health outcome questionnaires, particularly when used in association with hip or knee arthroplasty, and women tend to report greater pain and physical function limitation after hip or knee arthroplasty (Katz et al. 1994). 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 (Brinker et al. 1997, Hawker et al. 1998)]. 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 (Charnley 1979). Gender, age, and co-morbidity should be factored when comparing outcomes after hip or knee arthroplasty. Socioeconomic status probably does not have as significant an impact in a homogeneous country such as Sweden.

The mode of administration also significantly biases the results of health outcomes questionnaires. When a questionnaire is self-completed by the patient after knee surgery, as opposed to being administered by the investigator, the resulting questionnaire scores have been shown to be significantly worse (Hoher et al. 1997). Also, 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 (Hill et al. 1997). Therefore, an assessment of the status of non-responders is probably warranted when low response rate occurs with the administration of a questionnaire.

Selecting appropriate questionnaires

Since full and formal questionnaire validation was beyond the scope of this work, a questionnaire advocated for the SKAR should, at the very least, have undergone the validation process and have subsequently been deemed "valid". Many outcome questionnaires used for knee arthroplasty have not met this minimal standard. For those that have, not all have been validated specifically on the relevant arthroplasty population. Patients having undergone knee arthroplasty are older than the average population and are cardiovascularly fitter than age matched cohorts (Ries et al. 1996, Schroder et al. 1998). Therefore, it can not be automatically assumed that previously validated questionnaires will remain valid for use with this specific population. Questionnaires that are proposed for application to the SKAR should therefore be tested on the target population prior to wide-scale use.

The last decade has seen an increasing emphasis placed on determining the outcomes of prescribed medical/surgical interventions, and this is reflected in the large variety of outcome measures advocated in the literature. This holds true for the discipline of Orthopaedic Surgery. Unfortunately, there is scant consensus with respect to which outcome measures are most appropriate, and each author advocates their outcome measure over others using, at best, statistical methodology that makes direct comparison of measures difficult to interpret from a clinically useful vantage. Furthermore, while some measures are compared on homogeneous cohorts, most often the reader is forced to compare the value of a specific outcome questionnaire as contrasted with other questionnaires that have been tested on dissimilar patient populations. The problem is compounded by the constant introduction of new outcome measures, as opposed to focusing on those that exist. According to Streiner and Norman, "…perhaps the most common error committed by clinical researchers is to dismiss existing scales too lightly, and embark on the development of a new instrument with an unjustifiably optimistic and naïve expectation that they can do better" (Streiner et al. 1998). With this in mind, one of the aims of this research was to investigate existing questionnaires without advocating yet another new questionnaire. The characterization of a

more comprehensive endpoint other than revision status for knee arthroplasty appears to be possible with the use of existing health outcome questionnaires (Ritter et al. 1995, Hilding et al. 1997, Dawson et al. 1998, Hawker et al. 1998).

Broadly speaking, there are several categories of health outcome questionnaires that can range from a single item to hundreds of items that are summarized into multiple domains and summary scores. The categories include general health, disease specific, site specific, patient specific and single-item global questionnaires. General health questionnaires inquire about various aspects of patients' perception of their own health, including such diverse domains as ability to sleep, energy level, mood, and perception of body pain. General health questionnaires are not necessarily limited to any particular disease state nor patient cohort. The Nottingham Health Profile (NHP), 12-Item Short-Form Health Survey (SF-12), 36-Item Short-Form Health Survey (SF-36) and the Sickness Impact Profile (SIP) are examples of general health questionnaires. Disease specific questionnaires attempt to isolate the signal of interest by focusing questions around a particular disease state. The Western Ontario and MacMaster Universities Osteoarthritis Index (WOMAC) is an example. 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 Score is an example. Patient specific questionnaires 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 then asking them to rate or score how well those objectives have been accomplished. The Patient Specific Index is an example. Global, or single item, questionnaires are the most aggressive in their effort to limit noise by asking a single direct question regarding the state or condition of interest. Expanded definitions of each of these types of questionnaires are listed in the Methods section. Which categories of questionnaires to employ with the SKAR is unclear, but several authors have suggested that the simultaneous use of general health and disease/site specific questionnaires seems to yield complimentary data (Patrick et al. 1989, Hawker et al. 1995, Lieberman et al. 1997). This complimentary relationship speaks to the WHO definition of health and the consideration of mind and body as one.

Although there appears to be a vague consensus as to which categories of outcome questionnaires to apply to knee arthroplasty patients, there is no consensus whatsoever regarding specifically which questionnaires to use. Instead, a multitude of questionnaires have been put forward in the literature and new questionnaires continue to be introduced. Perspective researchers are forced subsequently to choose a questionnaire based on its published psychometric properties, or, perhaps more alarmingly, based on precedence and extraneous political factors. Choosing a questionnaire from the literature based on its psychometric properties is problematic.

Aims of the study

The aims of the study were as follows:

1. To investigate the feasibility of a large-scale postal survey of health outcome questionnaires to patients registered with the Swedish Knee Arthroplasty Registry.

2. To determine which general health and disease/site specific questionnaires were most appropriate for a large-scale application to knee arthroplasty patients registered with the Swedish Knee Arthroplasty Registry.

3. To investigate differences in feasibility and psychometric parameters between a global single-item outcome questionnaire and more comprehensive multi-item outcome questionnaires when assessing outcomes after knee arthroplasty.

4. To determine what patients are referring to when describing their level of satisfaction after knee arthroplasty.

5. To investigate what factors bias outcome questionnaires after knee arthroplasty.

6. To translate and validate the Oxford-12 Item Knee Score for use in Sweden

7. To determine the post-operative disposition of knee arthroplasty patients based on their pre-operative WOMAC scores, and to determine the sensitivity of specific items within the WOMAC to detect changes from pre and post-operative status.

Patients and methods

Literature review

In the winter of 1998, the National Library of Medicine Medline database was searched using the keywords "questionnaires" and "outcomes" in an effort to identify potential health questionnaires. Only those questionnaires that could generally be applied to Orthopaedic populations were selected and these included general health and disease/site specific measures. The disease/site specific measures unrelated to arthritis of the knee were not included.

Once a list of questionnaires were compiled, a further literature review using the same database was applied to the list looking in particular for references to five modifying criteria. These included 1) application of the questionnaire to knee arthroplasty patients, 2) application to patients with osteoarthritis, 3) previous validation studies, 4) use of the questionnaire in a postal survey format, and 5) translation and validation of the instrument into Swedish.

12 outcome measures were identified as potential candidates for further study—9 general health and 3 disease specific (Table 1). 5 of the general health questionnaires were excluded from further study as they had limited representation in the literature with respect to osteoarthritis and more specifically, arthroplasty. These included the COOP/WONCA, EuroQol, Functional Status Index, Index of Well Being, Duke-17, and the Musculoskeletal Functional Assessment (Table 1).

The remaining general health questionnaires, with the exception of the SF-12, all had precedence for application to osteoarthritis and arthroplasty patients, had all been translated into Swedish, had all been shown suitable for postal surveys, and all had their validity, reliability, and responsiveness previously determined (Table 1).

3 disease/site specific outcome measures were selected for further study, despite the large number identified in the literature (Drake et al. 1994, Sun et al. 1997). The principle reason that the majority of disease/site specific questionnaires related to the knee were excluded was that the majority relied on "objective" input from the surgeon and subsequently where not appropriate for the postal-survey mandate of further studies. The 3 disease/site specific questionnaires selected were the WOMAC, Oxford-12 Item, and the Lequesne Algofunctional. All 3 were relevant to osteoarthritis of the knee, and all 3 were valid, reliable and responsive (Table 1). The Oxford-12 had not been used in Sweden.

The SF-12 was selected for further study, despite its failure to meet the predefined criteria. The rationale for selecting the SF-12 was based on the fact that it is a select 12 of the 36 questions in the original SF-36, which had been widely applied to this patient population and is perhaps the most extensively validated and applied questionnaire. Furthermore, one of the underlying hypothesis of proposed studies was that the simpler a questionnaire is, then the greater the rate of compliance and efficiency of return from a postal version. Contrasting the SF-12 to the SF-36 would allow for direct investigation of this hypothesis.

There were few disease/site specific questionnaires for knee pathology that do not rely on the "objective" input of a clinical rater, usually the surgeon. Obviously, such questionnaires were not suitable for a postal survey and could be automatically eliminated from further consideration. This left few disease specific questionnaires for investigation, namely the WOMAC, Oxford-12 and Lequesne. The Lequesne is an established questionnaire which has been compared to the WOMAC in a double blind clinical trial (Bellamy et al. 1992). Furthermore, the Osteoarthritis Research Society and the 5th WHO/ILAR Task Force have advocated both the Lequesne and WOMAC as important outcome measures (Bellamy 1995). The Lequesne and WOMAC have both been used in Sweden.

The Oxford-12 item Knee Score was a new outcome measure derived from the Oxford-12 item Hip Score (Dawson et al. 1996b). This question-

Table 1. Studies listed by reference number previously demonstrating satisfactory fulfillment of each criteria for a given questionnaire

Questionnaire	Knee arthro-plasty	Osteoarthritis	Validation studies	Use in postal survey	Swedish translation
COOP/ WONCA	None identified	None identified	Kinnersley et al. 1994 McHorney et al. 1992	Essink-Bot et al. 1997	None identified
Duke-UNC/ Duke-17	None identified	None identified	Kaplan et al. 1976 Liang et al. 1990	None identified	None identified
EuroQol	None identified	None identified	Brazier et al. 1993 Hurst et al. 1997	Brazier et al. 1993 Dolan et al. 1996 Essink-Bot et al. 1997 Wolfe et al. 1997	None identified
FSI	Liang et al. 1990	Liang et al. 1990	Jette et al. 1986 Jette 1987 Liang et al. 1990	Liang et al. 1990	None identified
IWB	Liang et al. 1990	Liang et al. 1990	Jette et al. 1986 Jette 1987 Liang et al. 1990	Liang et al. 1990	None identified
MFA	None identified	Martin et al. 1997	Engelberg et al. 1996 Martin et al. 1996 Martin et al. 1997	Martin et al. 1997	None identified
NHP	Rissanen et al. 1995 Hilding et al. 1997	Hunt et al. 1981b Wiklund et al. 1988 Wiklund et al. 1991 Nilsson et al. 1994 Lescoe-Long et al. 1996 Franzen et al. 1997 Hilding et al. 1997	Hunt et al. 1980 Essink-Bot et al. 1997	Hunt et al. 1981b Wiklund et al. 1991 Lescoe-Long et al. 1996 Plant et al. 1996 MacDonagh et al. 1997	Wiklund et al. 1988 Wiklund et al. 1990 Wiklund et al. 1991
SF-12	None identified	Di Fabio et al. 1998	Ware et al. 1996 Jenkinson et al. 1997 Gandek et al. 1998)	None identified	Gandek et al. 1998
SF-36	Bombardier et al. 1995 Hawker et al. 1995 Williams et al. 1997	Bombardier et al. 1995 Hawker et al. 1995 Braeken et al. 1997 Williams et al. 1997	Brazier et al. 1992 McHorney et al. 1992 Jenkinson et al. 1994	Sullivan 1994 Sullivan et al. 1995	Sullivan 1994 Sullivan et al. 1995
SIP	Liang et al. 1990	Bergner et al. 1981 Laupacis et al. 1993 Stucki et al. 1995	Bergner et al. 1981 Deyo et al. 1983	Sullivan 1985 Sullivan et al. 1986	Sullivan 1985 Sullivan et al. 1986
Lequesne	Ryd et al. 1997	Lequesne et al. 1987 Lequesne et al. 1991 Bellamy et al. 1992 Lohmander et al. 1996 Lequesne et al. 1997	Lequesne et al. 1987 Lequesne 1989	None identified	Lohmander et al. 1996 Ryd et al. 1997 Translated but not validated
Oxford-12	Dawson et al. 1998	Dawson et al. 1996a Dawson et al. 1996b Dawson et al. 1998	Dawson et al. 1998	None identified	None identified
WOMAC	Bombardier et al. 1995 Hawker et al. 1995 Anderson et al. 1996 Williams et al. 1997	Bellamy et al. 1988 Bellamy 1989 Bellamy et al. 1991 Bellamy et al. 1992 Laupacis et al. 1993 Bombardier et al. 1995	Bellamy et al. 1988 Bellamy et al. 1992 Roos et al. 1998	Hawker et al. 1995	Roos et al. 1998

FSI = Functional Status Index
IWB = Index of Well-Being
MFA = Musculoskeletal Functional Assessment

naire had been applied to knee arthroplasty and osteoarthritis patients and had been shown to be valid, reliable and responsive (Dawson et al. 1998). However, it had not been used in Sweden previously. Still, the Oxford-12 Item Knee Score is simplistic enough in its question format without any particular cultural reference so that a rapid translation to Swedish would be sufficient to allow for further testing (Mathias et al. 1994).

Questionnaires (General Health)

Nottingham Health Profile (NHP) (Hunt et al. 1980, Hunt et al. 1981a, Wiklund et al. 1988)

The NHP poses 45 questions organized into 2 parts to which a response of yes or no is given. In Part 1, 38 questions are utilized to generate weighted scores for 6 domains, while in Part 2, 7 non-weighted questions are generated regarding perceived health problems affecting activities of daily life. Part 2 was not utilized in this study. Scores in Part 1 range from 0–100 with 0 representing the best possible health state. The domains for Part 1 are as follows: Pain, Physical Mobility, Energy, Emotional Reaction, Sleep, and Social Isolation

12-Item Short-Form Health Survey (SF-12) (Ware et al. 1996)

The SF-12 consists of 12 questions with Likert-box response key. Item scaling is both dichotomous and polychotomous. Scores are transformed into 2 weighted summary scores called Physical Component Summary and Mental Component Summary. The weights are calculated via a z and t-transformation so that an average population sample will record a score of 50 for each summary and a score change of 10 points represents one standard deviation. A score above 50 represents a perception of better health than the average population. For comparative purposes to other questionnaires, the SF-12 scores have been inverted in this study so that a score above 50 represents a perception of worse health than compared to an average population.

36-Item Short-Form Health Survey (SF-36) (Brazier et al. 1992, Ware et al. 1992, Sullivan et al. 1995)

The SF-36 consists of 36 questions with Likert-box response keys. Item scaling is both dichotomous and polychotomous. 8 domains scores are generated ranging from 0–100. The 8 domains are as follows: Body Pain, Physical Functioning, Vitality, General Health, Social Functioning, Role-Physical, Role-Emotion, and Mental Health. A score of 100 represents the best possible health state. 2 summary scales are also generated for the SF-36 (Physical and Mental Component Summary) and their scoring is similar as for the summary scores of the SF-12. Like the SF-12, the scores for the SF-36 have been inverted for comparative purposes.

Sickness Impact Profile (SIP) (Pollard et al. 1976, Sullivan 1985)

The SIP is a 136-item questionnaire that calls on patients to affirm a question with a simple check mark if it applies. Otherwise, the question response key is left blank (Damiano 1996). The questionnaire produces weighted results for 12 domains as well as 3 summary scores. The domains of the SIP include Body Care and Movement, Ambulation, Home Management, Mobility, Sleep and Rest, Alertness Behaviour, Recreation and Pastimes, Social Interaction, Emotional Behaviour, Communication, Work, and Eating. The summary scores include a Physical Dimension, a Psychosocial Dimension, and a Total Score. Scores range from 0–100 with 0 representing the best possible health state.

Questionnaires (Disease Specific)

Lequesne Index of Severity-Knee (Lequesne) (Lequesne et al. 1987, Lequesne 1997b)

The Lequesne consists of 11 questions with various scales utilized for different questions. Questions refer to Pain (5 questions), Walking (2 questions) and Activities of Daily Living (4 questions). Weights are applied in the scoring algorithm and a score range from 0 to 24 is produced. A score of 0 represents a perfect health state.

Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) (Bellamy et al. 1988, Roos et al. 1998)

The WOMAC consists of 24 Likert-box questions broken down into 3 domains: Pain (5 questions), Stiffness (2 questions) and Physical Function (17 questions). Scores range from 0-20 for Pain, 0-8 for Stiffness and 0-68 for Physical Function. A score of 0 represents the best possible health state. The items are scaled with five boxes for each question ranging from 0 to 4.

Questionnaires (Joint Specific)

Oxford-12 Item Knee Score (Oxford-12) (Dawson et al. 1998)

12 questions are posed relating specifically to the knee. Each question has a Likert-box response key from 1 to 5. A single score is produced ranging from 12 to 60, with 12 indicating the best possible health state.

Questionnaires (Single-Item Global Scores)

Satisfaction Questionnaire

A single-item questionnaire was employed using Likert-type boxes over a 4-point scale. Patients were asked specifically if they were satisfied with their knee arthroplasty. The 4 possible responses were 1) very satisfied 2) satisfied 3) uncertain or 4) unsatisfied. This questionnaire is unique to the SKAR and has not been previously validated.

Single-Item Knee Questionnaire

In an effort to avoid possible confounding noise from a multitude of items within a disease or joint specific questionnaire, a single-item questionnaire was developed for use with the SKAR. The question posed was as follows: On a scale from 1 to 10, how would you rate the result of your knee arthroplasty (1 being the best possible result and 10 being the worst possible result).

Single-Item General Health Questionnaire

Like the single-item knee score, a single-item questionnaire on general health was developed for use with the SKAR. The question posed was as follows: On a scale from 1 to 10, how would you rate your overall health (1 being the best possible and 10 being the worst possible result).

Questionnaires (Co-morbidity)

Modified Charnley Class for Knee Arthroplasty

Charnley proposed a co-morbidity scale when assessing the outcomes after total hip arthroplasty in 1979 (Charnley 1979). This rating scale used 4 graduated classes for co-morbidity ranging from monoarticular hip arthroplasty (Charnley A), monoarticular hip arthroplasty with contralateral hip osteoarthritis (Charnley B), bilateral hip arthroplasty (Charnley BB), and a systemic medical condition or remote osteoarthritis (e.g. knees, spine, etc) that impaired locomotory ability (Charnley C). Once a patient progressed from 1 category to the next, such as from Charnley B to C, they always remained in the worse category. That is, a change in Charnley class is unidirectional.

For the purposes of this study, the Charnley class was modified as follows: monoarticular knee arthroplasty (Charnley A), monoarticular knee arthroplasty with contralateral knee osteoarthritis (Charnley B1), bilateral knee arthroplasty (Charnley B2), and a systemic medical condition or remote osteoarthritis (e.g. hips, spine, etc) that impaired locomotory ability (Charnley C). Charnley B and BB were changed to B1 and B2 in order to facilitate easier computer based data searches, as a search for Charnley B would otherwise yield all B's and BB's.

All patients by definition had at least one knee arthroplasty in-situ as they were registered with the SKAR and therefore by default were considered Charnley A. Patients, as mentioned above, who had bilateral knee arthroplasties had one knee (left or right) randomly selected for the purpose of inquiry. The modified Charnley Class was determined using a 4-item questionnaire. The questions posed were as follows: 1) Do you have arthritis in your other knee (Charley B1), 2) do you have an artificial knee joint in your other knee (Charnley B2), 3) do you have arthritis in other joints besides

your knees, for example, your hips, feet or spine, that limits your ability to walk (Charnley C) and 4) do you have a medical condition that limits your ability to walk, for example, ischemic heart disease, congestive heart failure, emphysema, etc. (Charnley C).

Questionnaires (Patient Burden)

In order to determine the burden imposed on questionnaire respondents, a simple questionnaire was developed. Patients were asked to record the time, in minutes, that they required to complete a particular questionnaire and to record if they required assistance in order to complete the questionnaire (yes or no).

Feasibility

Questionnaire feasibility was investigated by multiplying the return rate of a questionnaire by the percentage of those questionnaires returned which were complete with responses for all items. Imputation was not used for missing items.

Translation into Swedish

It is insufficient to simply translate a questionnaire into another language (Guillemin et al. 1993, Guyatt 1993). Therefore, an effort was made in this thesis to use questionnaires that had previously been translated into Swedish. The only questionnaire employed that had not been previously translated and validated in Swedish was the Oxford-12, and its translation and validation forms part of this thesis (Paper V). The translation processes followed general guidelines from the literature (Guillemin et al. 1993, Mathias et al. 1994). The Oxford-12 was independently translated into Swedish and back translated by 1 professional translator and 1 bilingual Orthopaedic surgeon. A bilingual panel assessed adequacy of the translated versions and a final translated version was agreed upon. A pilot study was conducted on 8 bilingual subjects who completed in random order the Swedish and English version of the Oxford-12, separated by a 5-day interval, to further asses the translation.

Demographics recorded by the SKAR

Personal identification number

All citizens of Sweden receive a unique personal identification number (PIN) that is supplied and followed by the National Census Register (NCR). The PIN contains information regarding a person's date of birth and must be presented upon any encounter with government agencies, including hospitals. Ultimately, the PIN is linked to date of death. Because of the pervasiveness and acceptance of the PIN, knee arthroplasty patients, for example, are able to be comprehensively followed with regards to address change and initial and repeat encounters with the health care system up to and including date of death. This has made the Swedish National registries possible and the lack of such a cohesive number is an obstacle to comprehensive outcome registries in North America. Other Scandinavian countries also use a PIN equivalent.

PIN, knee arthroplasty, and side operated on

The SKAR records the PIN for each patient that undergoes knee arthroplasty surgery. A letter representing left or right side is added to the PIN so that each knee arthroplasty has a unique identification number. Subsequently, reports from the SKAR often contain reports of x number of knees operated on for a given period in y number of patients. The number of knees operated on is obviously larger than the number of patients, as some patients have bilateral knee arthroplasties.

Patient selection

Papers I

All knees operated on from 1981 to 1995 were identified and the associated PIN was cross-referenced to the NCR. This allowed for the identification of 28,962 unique knees operated on over this period in patients that were not recorded as deceased. Of the 28,962 knees operated on during

1981–1995, the postal office could not locate 122 and 133 envelopes were returned because the patient was said to be too ill or infirm to answer. The question on satisfaction was answered for 27,372 knees (95%), and these were the basis for the analyses. 22,866 (83.5%) knees had been operated for osteoarthrosis, 3,490 (12.8%) for rheumatoid arthritis, 515 (1.9%) for posttraumatic disorders and 206 (0.8%) for osteonecrosis. Various conditions accounted for the remaining 295 knees (1.0%). The average follow-up period was 6 (2–17) years after primary arthroplasty

Papers II, III, and IV

9 months after the postal survey in Paper I, 3,600 knees were randomly selected from the 27,372 knees selected for Paper I. A patient with bilateral knee arthroplasties had an equal chance of the left or right knee selected, however, once a side had been selected for a patient, the patient was removed from the eligible pool so that patients with bilateral knee arthroplasties would only receive 1 questionnaire package. Therefore, in this aspect of the thesis, number of knees equals number of patients. The random sample was restricted to patients with a diagnosis of primary osteoarthritis, age ≥ 55 at time of surgery, age ≤ 95 at the time of mail-out and prosthesis type of medial uni-compartmental, lateral uni-compartmental, bilateral (same knee) uni-compartmental and total knee arthroplasty. Patients who were registered as having undergone a revision were eligible, providing they were not known to have had an extraction arthroplasty, amputation or arthrodesis.

The 3,600 selected patients were randomly divided into 12 groups of 300, each receiving a combination of 1 general health and 1 disease/site specific questionnaire (4 general health questionnaires x 3 disease/site specific questionnaires). All patients received a cover letter with instructions and a postage-paid return envelope, a 3rd questionnaire regarding co-morbidity (Co-morbidity Questionnaire, described above), a 4th questionnaire inquiring about the length of time required and the need for assistance to complete the questionnaires (patient Burden Questionnaire, described above), and a 5th questionnaire regarding satisfaction (Satisfaction Questionnaire, described above). The Satisfaction Questionnaire was the same as for in Paper I. A reminder letter was sent at 2 weeks for non-responders.

The average patient age at the time of mail-out was 78 (57–94) years and 71 (55–90) years at the time of index surgery. The average follow-up time was 7 (1–23) years. 69.8% (n=2511) of the sample were women and 30.2% (n=1089) were men. 94.5% had not undergone revision surgery (removal, addition or exchange of a component). 57.9% had tri-compartmental knee replacements, 36.0% had medial uni-compartmental knee replacements, leaving 6.1% with either a lateral uni-compartmental or both compartments of the same knee replaced with a uni-compartmental prosthesis.

Paper V

A subset of 1200 of the patients (knees) from Papers II, III, and IV were analyzed in this paper. The 1200 patients were from the 4 groups of 300, each receiving a combination of 1 of 4 general health questionnaires along with the Oxford-12. As in Papers II, III, and IV, all patients received a cover letter with instructions and a postage-paid return envelope, a 3rd questionnaire inquiring about the length of time required and the need for assistance to complete the questionnaires, and a 4th questionnaire regarding satisfaction. A reminder letter was sent at 2 weeks for non-responders. At 3 weeks, 120 patients were randomly selected from those that completed the Oxford-12 and were sent a WOMAC.

The average patient age at the time of mail-out was 78 (58–94) years and 71 (55–90) years at the time of index surgery. The average follow-up time was 7 (1–21) years. 70% (n=840) of the sample were women and 30% (n=360) were men. 94% were primary arthroplasties. 59% of all patients had tri-compartmental knee replacements, 35% had medial uni-compartmental knee replacements, and 6.0% had either a lateral uni-compartmental or both compartments of the same knee replaced with an uni-compartmental prosthesis.

Paper VI

156 primary total knee arthroplasties with a diagnosis of osteoarthritis operated on from period November 1995 to April 1998 were followed prospectively in a multi-centre Canadian trial. The

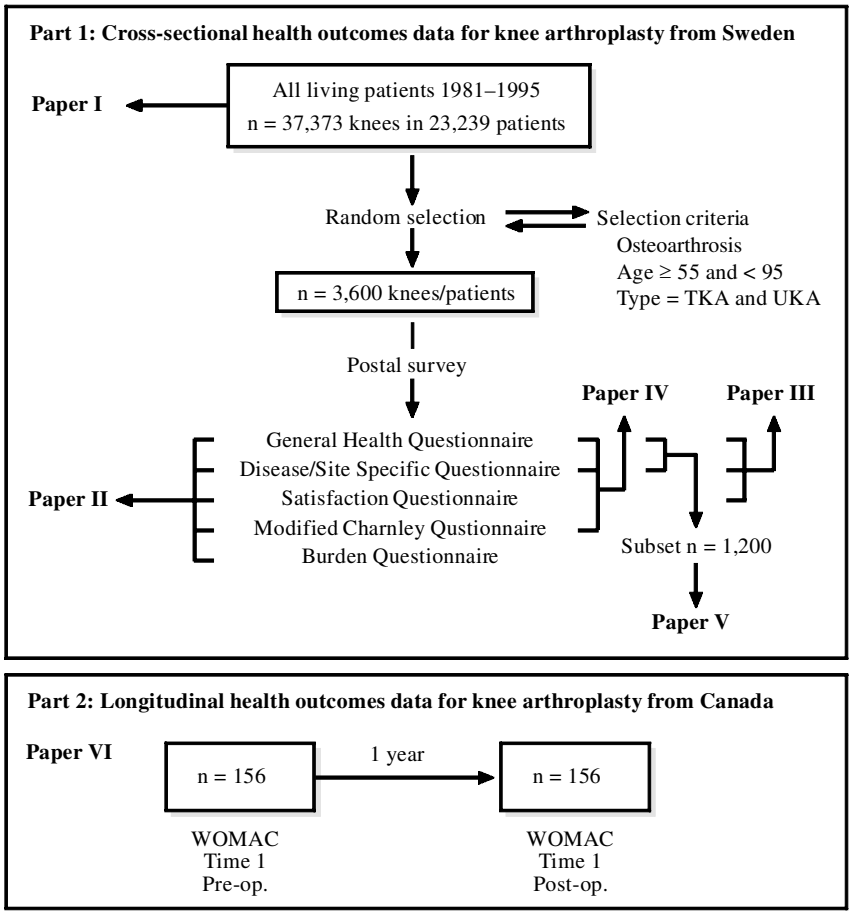

Figure 5. Schematic representation of patient selection and breakdown for Papers contained in this thesis.

average patient age at the time of surgery was 75 (50–92) years. 53% (n=83) were women. 149 Genesis and 7 Genesis II prosthesis were inserted in 156 patients using a paramedial arthrotomy. 96% (n=149) had a patellar resurfacing and the PCL was preserved in all cases. All patients completed a WOMAC preoperatively and at 1-year postoperatively.

An overview of the patient selection for this thesis appears in Figure 5.

Statistics

For all tests in which a P-value has been calculated, $P < 0.05$ has been considered as significant. 95% confidence intervals have been supplied where appropriate. The Chi Squared test has been used to investigate differences in frequency distributions of data. Parametric tests (Student's t-test, ANOVA) have been used with continuous data, such as time required to complete a questionnaire, while non-parametric tests (Mann Whitney U-test, Kruskal Wallis test) have been used with the ordinal data produced by questionnaires. Multinomial regression was performed to determine the variables that significantly affected the modified Charnley Class. Multilinear regression was performed to determine the variables that significantly affected each specific questionnaire.

The non-parametric Spearman's correlation coefficient was used when correlating the results of a questionnaire against a construct. The intraclass correlation coefficient was used for test-retest reliability and Cronbach's alpha statistic was used to investigate internal consistency reliability. Values

of 0.6–0.8 for these 2 tests have been defined as fair, 0.8–0.9 as good, and >0.9 as excellent. For single-item questionnaires, the weighted Kappa coefficient has been used for test-retest reliability. A Kappa coefficient of 0.4–0.6 was defined as fair, 0.6–0.8 good, and >0.8 excellent. Responsiveness was indirectly assessed using the ROC Curve method, with an area under the curve of 0.5 defined as a non-discriminating test and an area of 1.0 as a perfectly discriminating test.

SPSS ® Version 9.0 software was used for all calculations other than the weighted Kappa for which Analys-It ® was used.

Ethics approval

For research conducted in Sweden (Papers I–V), comprehensive permission from the Swedish Health Authority (Socialstyrelsen) and the National Controlling Body for Computer Registries (Datainspektionen) was granted to obtain and record patient factors related to knee arthroplasty. For research conducted in Canada (Paper VI), ethics approval was obtained from the Ethical review Boards of the participating university hospitals.

Summary of Papers

Paper I: Patient satisfaction after knee arthroplasty. A report on 27,372 knees operated on between 1981 and 1995 in Sweden

Introduction

The validation of the SKAR afforded an opportunity to inquire about patient satisfaction regarding their knee arthroplasty. However, to avoid a potential reduction in response rate to the critical validation questionnaire, an inquiry about satisfaction needed to be short and simple. A single-item Likert-type questionnaire regarding satisfaction was developed. Patients were asked to affirm 1 of a continuum of 4 possible responses, indicating how satisfied they were with the operated knee. The possible responses were as follows: 1) very satisfied 2) satisfied 3) uncertain or 4) unsatisfied.

Methods

28,962 living patients identified were mailed a 2-part questionnaire regarding the revision status of their knee along with the single-item satisfaction questionnaire. A reminder letter was sent at 4 weeks for non-responders. As the satisfaction questionnaire was single-item, missing responses could not be imputed. The question on satisfaction was answered for 27,372 knees (95%), and these are the basis for the analyses. The questionnaire regarding revision was used in a validation study of the SKAR (Robertsson et al. 1999b).

Answers were classified on an ordinal scale (unsatisfied < uncertain < satisfied < very satisfied) and compared and evaluated for different selections of patients. When comparing age differences between sexes, Student's t-test was used. Non-parametric analyses (Mann Whitney U-test and Kruskal Wallis H-test) were used when comparing satisfaction between groups. For correlation, the non-parametric Spearman correlation coefficient was used.

Results

27,372 (95%) patients operated on between 1981 and 1995 responded. Of those responding 81% were satisfied or very satisfied, 11% uncertain and 8% were unsatisfied. The proportion of satisfied patients was affected by the pre-operative diagnosis, with patients with rheumatoid arthritis being the most satisfied, followed by patients operated for osteoarthrosis, post-traumatic condition and osteonecrosis (Kruskal Wallis, $p<0.001$) (Figure 6). There was no difference in the proportional distribution of satisfaction status between patient groups operated on with a TKA, a medial UKA, or a lateral UKA (Figure 7). Bilateral (same knee) UKA, however, had a significantly higher proportion of dissatisfied patients (Kruskal Wallis, $p=0.04$). Patellar resurfacing in primary TKA yielded a higher ratio of satisfied patients than for unresurfaced patellae, but this increased ratio diminished with time passed since the primary operation. In unrevised cases the overall satisfaction rate was unchanged regardless of the time passed

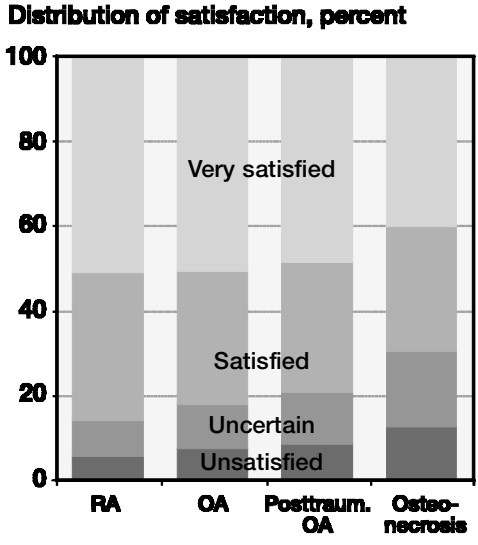

Figure 6. In unrevised cases, 14% of 3,203 RA patients, 18% of 21,165 OA patients, 2% of 449 postraumatic OA patients and 30% of 191 patients with osteonecrosis were unsatisfied or uncertain.

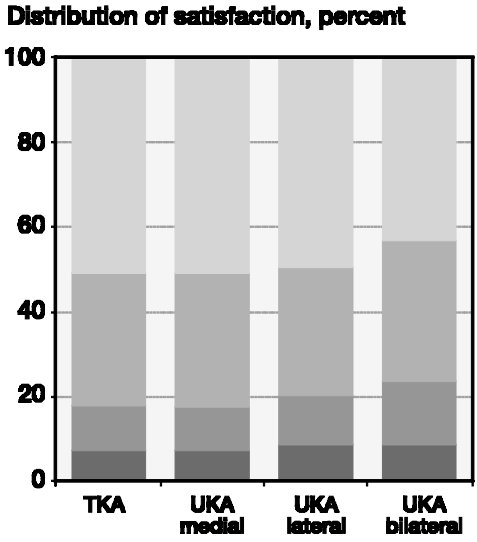

Figure 7. In unrevised OA cases, 18% of 12,298 TKAs, 17% of 7,860 medial UKAs, 20% of 686 lateral UKAs and 23% of 150 medial + lateral UKAs were unsatisfied or uncertain.

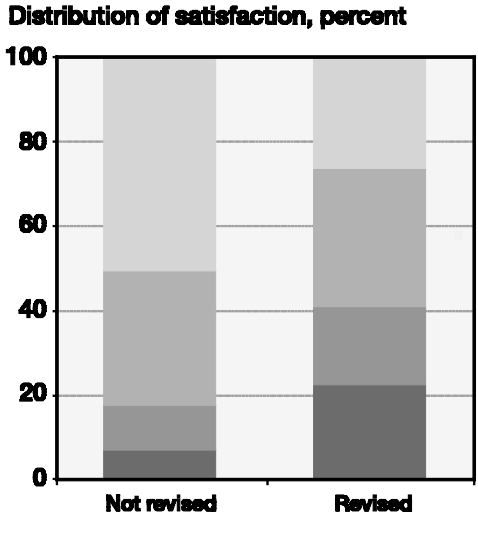

Figure 8. 17% of 25,275 unrevised cases (all types and diagnoses) and 41% of 2,097 revised cases were unsatisfied or uncertain.

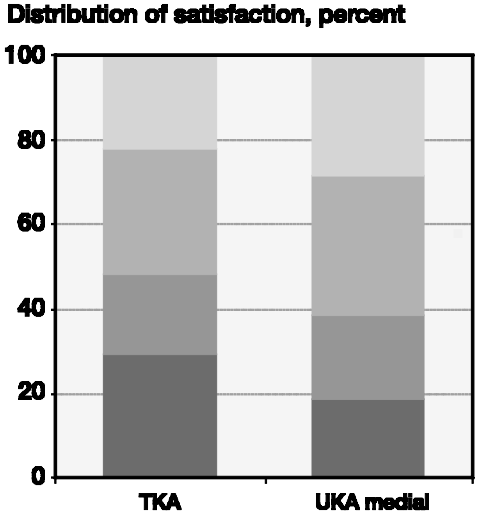

Figure 9. In OA, 48% of 668 revised TKAs and 39% of 887 revised medial UKAs were unsatisfied or uncertain.

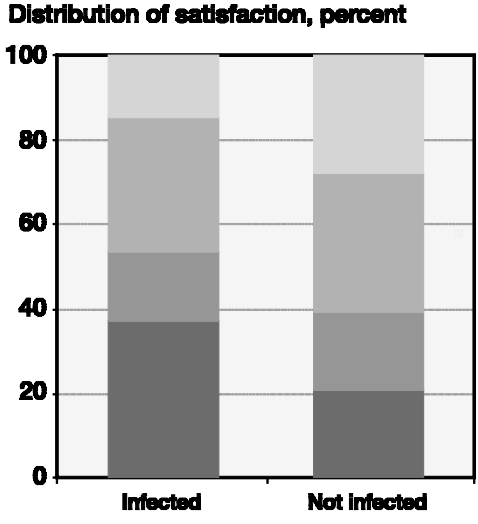

Figure 10. In revised cases, 53% of 232 who were revised for infection and 39% of 1,865 who were revised for other reasons were unsatisfied or uncertain.

since the primary operation. The proportion of satisfied patients was higher in unrevised knees than in revised knees in which 22% of patients were unsatisfied after a mean follow-up of 5 (0-16) years (Mann Whitney, $p<0.001$) (Figure 8). Revised UKA had a higher proportion of satisfied patients than a revised TKA (Mann Whitney, $p<0.001$) (Figure 9). Revision for infection yielded a higher ratio of unsatisfied patients than for revision for other reasons (Mann Whitney, $p<0.001$) (Figure 10).

Conclusions

A simple satisfaction questionnaire has an exceptionally high response rate, for this population, and can generate useful comparative outcomes

data. For example, the chronicity of onset of pathology leading to knee arthroplasty directly correlates to post-operative patient satisfaction. Large proportions of knee arthroplasty patients are satisfied with the intervention, even after revision, for non-infected reasons. Infection has a profound effect on patient satisfaction. Successful knee arthroplasty can be expected to result in long-lasting patient satisfaction.

Paper II: Appropriate questionnaires for knee arthroplasty: Results of a survey to 3600 patients from the Swedish Knee Arthroplasty Registry

Introduction

The only outcome metric available for use with the SKAR has been revision status. While definitive and precise, revision status yields data on the small minority of operations that fail but tells us nothing of the status of the majority of patients. Health outcome questionnaires can be used to define more comprehensive endpoints. Numerous questionnaires are available for application to a knee arthroplasty population, but there is no consensus as to which are the most appropriate to use. Reaching a consensus is confounded by the fact that there are no gold standards by which to judge questionnaires for knee arthroplasty.

Direct comparison of questionnaires for knee arthroplasty is not possible with the published literature, as the psychometric properties of many of the questionnaires advocated have not been determined. Of those that have, the properties have often been determined on a general population. Knee arthroplasty patients are distinct from an aged matched general population in that they are fitter and have a longer life expectancy. Previously defined properties of questionnaires may therefore not be directly transferable to this unique population.

The purpose of this study was to identify relevant general health and disease/site specific outcome questionnaires for knee arthroplasty and simultaneously test them on a large random sample from the SKAR. It was hypothesized that differences in the validity and reliability properties as well as feasibility and patient burden would differ by questionnaire and that some would be more appropriate for this application than others.

Methods

4 general health questionnaires (NHP, SF-12, SF-36, and the SIP) and 3 disease/site specific questionnaires (Lequesne, Oxford-12 and the WOMAC) were sent in a postal survey to 3600 randomly selected patients from Paper I. Patients were randomly divided into 12 groups of 300, each receiving a combination of 1 general health and 1 disease/site specific questionnaire (4 general health questionnaires x 3 disease/site specific questionnaires).

3 weeks after the first mailing, 420 (60 patients x 7 questionnaires) patients were randomly selected from those that had responded to the first mail-out and were sent 1 repeat questionnaire (generic or disease/site specific) in order to test the reproducibility of each questionnaire.

Response rate, patient burden, content validity and reliability were calculated for each of the 4 general health and 3 disease/site specific questionnaires. Ranks were assigned for each of the tested parameters for each questionnaire. An average rank for each questionnaire by class (general health or disease/site specific) was generated.

Results

84.8% (n=3052) of patients responded by returning their questionnaire. The response rates for the SF-12, SF-36, and NHP (87.4%, 86.6%, and 85.3%, respectively), were significantly higher than for the SIP (81.4%, Chi-square $p < 0.001$). There was no difference in the response rates for the disease/site specific questionnaires.

For the general health questionnaires the SF-12 had the highest percentage of questionnaires returned completed (75.4%, Chi-square, $p < 0.001$). The SIP (67.9%) and the NHP (67.2%) were indistinct form each other. The SF-36 had a significantly lower efficiency of completion (63.0%, Chi-square, $p < 0.001$). The Oxford-12 had a significantly higher (Chi-square, $p < 0.001$) percentage of complete questionnaires for the disease/site specific questionnaires (89.4%) followed by the WOMAC (83.0%) and the Lequesne (79.1%) (Table 2).

Table 2. Gross and net response rates for general health and disease/site specific questionnaires

Questionnaire	Number received by patients [a]	Number returned	Gross % returned (95% CI [d])	% complete [b] (95% CI [d])	Net % return [c] (95% CI [d])
General health					
Nottingham Health Profile	896	764	85.3 (85.2–85.4)	67.2 (67.1–67.3)	57.3 (57.2–57.4)
SF-12	895	782	87.4 (87.3–87.5)	75.4 (75.3–75.5)	65.9 (65.8–66.0)
SF-36	899	779	86.6 (86.5–86.7)	63.0 (62.9–63.1)	54.6 (54.5–54.7)
Sickness Impact Profile	893	727	81.4 (81.3–81.5)	67.9 (67.8–68.0)	55.3 (55.2–55.4)
Disease specific					
Lequesne	1194	1012	84.8 (84.7–84.9)	79.1 (79.0–79.2)	59.2 (59.1–59.3)
Oxford-12	1194	1026	85.9 (85.8–86.0)	89.4 (89.3–89.5)	76.7 (76.6–76.8)
WOMAC	1195	1014	84.9 (84.8–85.0)	83.0 (82.9–83.1)	70.5 (70.4–70.6)

[a] Number of patients sent a questionnaire package minus those returned by post office or with note indicating that the patient was deceased.
[b] Percentage of questionnaires returned that were fully completed.
[c] *Percent net return* equals *percent returned* multiplied by *percentage complete*.
[d] 95% Confidence interval

The highest net percentage of completed general health questionnaires was for the SF-12 (65.9%) followed by the NHP (57.3%), SIP (55.3%) and the SF-36 (54.6%). The Oxford-12 was the highest for the disease/site specific questionnaires (76.7%) followed by the WOMAC (70.5%) and the Lequesne (59.2%).

The time required to complete all general health and disease/site specific questionnaires were significantly different (ANOVA, $p < 0.0001$). The SIP required the most time for completion (23 minutes) and the SF-12 the least (8 minutes). The WOMAC required the most time to complete for the disease/site specific questionnaires (12 minutes) and the Lequesne the least (8 minutes). Patients reported a significantly greater frequency (29%) of requiring assistance to complete the SF-36 as compared to the other general health questionnaires (Chi-square, $p = 0.005$). Similar frequencies for requiring assistance were observed for the disease/site specific questionnaires.

Considerable variation in floor and ceiling effects were seen between general health and disease/site specific questionnaires (Table 3). The average intraclass correlation coefficients for the general health questionnaire group ranged from 0.91 (NHP) to 0.75 (SF-36). The highest intraclass correlation coefficient for the disease/site specific questionnaires ranged from 0.94 (Oxford-12) to 0.85 (Lequesne). Cronbach's alpha coefficient for the SF-12 was lower than for all others (0.62).

The SF-12 ranked best overall for the general health questionnaires and the Oxford-12 ranked best overall for the disease/site specific questionnaires when the individual ranks for each parameter were averaged (Table 4).

Conclusions

Considerable variation was found in the performance of multiple questionnaires when measured by various standards. The SF-12 and Oxford-12, however, had the best overall ranking for a general health and disease/site specific questionnaire, respectively based on the tested criteria. These questionnaires can be considered the most appropriate for use in a wide-scale discriminative postal-survey to the SKAR. The Lequesne, WOMAC, SF-36 and NHP performed satisfactory. Based on poor performance over multiple parameters, the use of the SIP in this context can not be recommended. Questionnaires should be tested on the target population prior to wide-scale use.

Table 3. Breakdown of reliability and construct validity factors as well as scores by domains for general health and disease/site specific questionnaires

Questionnaire	Reliability		Content validity			Scores		
	Cronbach's alpha [a]	ICC [b]	Floor	Ceiling	Skew	Avgerage [c] (95% C.I)		Possible score range
GENERAL HEALTH								
Nottingham Health Profile (n=764)								
Emotional Reaction	0.85	0.84	58.37	1.16	2.04	13.3	(11.6–15.0)	0–100
Sleep	0.72	0.89	27.99	2.79	1.10	25.1	(23.1–27.1)	0–100
Energy	0.64	0.91	49.59	19.61	0.68	33.8	(30.9–36.7)	0–100
Pain	0.85	0.95	38.10	2.77	1.18	23.0	(20.9–25.0)	0–100
Physical Mobility	0.80	0.97	25.11	1.56	0.71	28.2	(26.4–30.1)	0–100
Social Isolation	0.60	0.87	74.97	0.42	2.37	9.3	(7.9–10.7)	0–100
Average	0.74	0.91	45.69	4.72	1.35	N/A		
SF-12 (n=782)								
Physical Component Summary	0.62	0.85	0.02	0.00	0.25	37.3	(36.4–38.1)	0–100
Mental Component Summary	0.62	0.92	0.02	0.00	-0.42	49.7	(48.8–50.7)	0–100
Average	0.62	0.88	0.02	0.00	-0.09	N/A		
SF-36 (n=779)								
Physical Functioning	0.90	0.89	0.79	5.83	-0.14	43.2	(41.2–45.2)	0–100
Role-Physical	0.88	0.57	21.32	49.50	-0.69	34.1	(31.1–37.0)	0–100
Body Pain	0.92	0.86	17.70	3.37	-0.14	56.3	(54.1–57.6)	0–100
General Health	0.81	0.88	3.26	0.59	0.02	55.9	(54.1–57.6)	0–100
Vitality	0.82	0.69	3.26	1.88	0.08	52.9	(51.0–54.8)	0–100
Social Functioning	0.75	0.77	36.02	2.45	0.85	73.5	(71.4–75.5)	0–100
Role-Emotion	0.88	0.71	41.41	36.08	0.09	52.4	(49.1–55.7)	0–100
Mental Health	0.83	0.80	12.82	0.43	0.68	72.1	(70.5–73.8)	0–100
Transition	N/A	0.56	N/A	N/A	0.21	3.2	(2.3–4.2)	0–100
Average	0.85	0.75	17.07	12.52	0.11	N/A		
Physical Component Summary	N/A	0.93	N/A	N/A	-0.29	33.3	(32.4–34.3)	0–100
Mental Component Summary	N/A	0.82	N/A	N/A	0.43	47.9	(46.8–49.1)	0–100
Sickness Impact Profile (n=727)								
Sleep and Rest	0.62	0.81	40.54	0.57	1.84	22.4	(21.2–23.6)	0–100
Emotional Behaviour	0.80	0.96	68.49	0.72	3.16	23.4	(22.2–24.6)	0–100
Body Care and Movement	0.88	0.87	42.67	0.58	2.37	18.5	(17.3–19.6)	0–100
Home Management	0.86	0.87	52.77	46.37	1.62	34.1	(32.3–35.8)	0–100
Mobility	0.81	0.89	59.97	0.88	2.59	23.0	(21.8–24.3)	0–100
Social Interaction	0.88	0.76	46.04	0.72	3.83	13.2	(12.3–14.2)	0–100
Ambulation	0.82	0.88	28.12	0.43	1.14	18.6	(17.3–20.0	0–100
Alertness Behaviour	0.85	0.63	67.91	1.16	3.12	25.5	(24.2– 26.8)	0–100
Communication	0.75	0.73	75.00	0.58	4.27	19.8	(18.9–20.6)	0–100
Work	N/C	0.68	69.00	0.23	1.14	61.7	(58.8–64.5)	0–100
Recreation and Pastimes	0.71	0.85	37.98	0.30	1.25	27.6	(26.1–29.1)	0–100
Eating	0.84	0.52	80.97	0.59	8.64	2.3	(1.6–2.9)	0–100
Average	0.80	0.79	55.79	4.43	2.91	N/A		
Physical Dimension	N/A	0.92	28.15	0.30	2.18	12.3	(11.1–13.4)	0-100
Psychosocial Dimension	N/A	0.87	41.63	0.44	4.07	6.8	(5.9–7.6)	0-100
Total Score	N/A	0.97	22.98	0.00	2.57	8.9	(7.9–9.9)	0-100
DISEASE/SITE SPECIFIC								
Lequesne (n=1012)	0.77	0.85	6.38	0.00	0.42	8.9	(8.6–9.3)	0–25
Oxford-12 (n=1026)	0.93	0.94	6.76	0.11	0.73	25.5	(24.9 –26.2)	12–60
WOMAC (n=1014)								
Pain	0.91	0.95	20.48	0.52	0.72	5.1	(4.8–5.4)	5–25
Stiffness	0.91	0.90	25.76	1.93	0.54	2.3	(2.2–2.4)	2–10
Physical Function	0.98	0.92	8.59	0.12	0.34	23.0	(21.9–24.2)	17–75
Average	0.93	0.92	18.27	0.85	0.53	N/A		

[a] See methods for description of Cronbach's Alpha.
[b] ICC = Intra-Class Correlation Coefficient – see methods for description.
[c] Geometric mean.

Table 4. Average ranked values for general health and disease/site specific questionnaires for each parameter (1 = highest rank, 4 = lowest rank)

Questionnaire	Burden		Feasability		Content validity			Reliability		Average rank
	Time	Help	Response	% Compl.	Floor	Ceiling	Skew	ICC [a]	Cr. alpha [b]	
General health										
Nottingham Health Profile	2	2	3	3	3	3	3	1	3	2.6
SF-12	1	3	1	1	1	1	1.5	2	4	1.7
SF-36	3	4	2	4	2	4	1.5	4	1	2.8
Sickness Impact Profile	4	1	4	2	4	2	4	3	2	2.9
Disease/site specific										
Lequesne	1	3	3	3	1	1	1	3	3	2.1
Oxford-12	2	1	1	1	2	2	3	1	1.5	1.6
WOMAC	3	2	2	2	3	3	2	2	1.5	2.3

[a] ICC = Intraclass Correlation Coefficient – see methods for description.
[b] Cr. alpha = Cronbach's alpha – see methods for description.

Paper III: Patient satisfaction compared with general health and disease specific questionnaires in 3600 patients operated on with knee arthroplasty

Introduction

Health outcome questionnaires can be cumbersome, and it is known that for self-administrated postal surveys that the higher the patient's burden, the lower the response rate. Thus, when evaluating questionnaires for use in postal surveys, not only does the usual psychometric properties have to be taken into account, but also the response rate and completeness. When studying a phenomenon with a low incidence or prevalence in the target population, a small loss in patient response rate may significantly effect the analysis. In such instances, or when extensive questionnaires can not be used for practical reasons, a single-item questionnaire on satisfaction might yield useful information regarding the effect of the intervention. Further, when the preoperative status has not been recorded, as was the case with the SKAR, patients can be assumed to take their pre-operative condition into account when answering and thus act as their own comparison. Partly for these reasons, a simple questionnaire on satisfaction was developed and sent to all living patients registered with the SKAR from 1981 to 1995 (Paper I).

To evaluate what knee arthroplasty patients are referring to when answering a question regarding satisfaction with the procedure, and to partially validate the questionnaire, the results of the patient satisfaction questionnaire were compared with the results of general health (NHP, SF-36, SF-12) and disease/site specific (Oxford-12, WOMAC) questionnaires.

Methods

In August of 1997 a postal survey was sent to all living patients (32,428 knees in 27,114 patients) registered with the SKAR as part of a validation study. A single-item questionnaire regarding satisfaction was included in this survey. 9 months later, in May 1998, a more elaborate study of health status was performed by a postal survey to 3600 randomly selected osteoarthrosis patients from the SKAR (Paper II). The patients were divided into random groups that were sent different combinations of health questionnaires so that each patient received 1 general health and 1 disease/site specific questionnaire along with the above mentioned single-item satisfaction questionnaire.

The reliability (Kappa coefficient) of the short satisfaction question was determined by comparing the August 1997 answers of previously unrevised patients with their answers from May 1998. Patients revised between the 2 postal surveys could be assumed to have a change in their knee condition and were excluded. This left 2711 patients that had answered on both occasions. Responsiveness of the satisfaction questionnaire was indirectly assessed using the ROC Curve method using revised and unrevised patients as

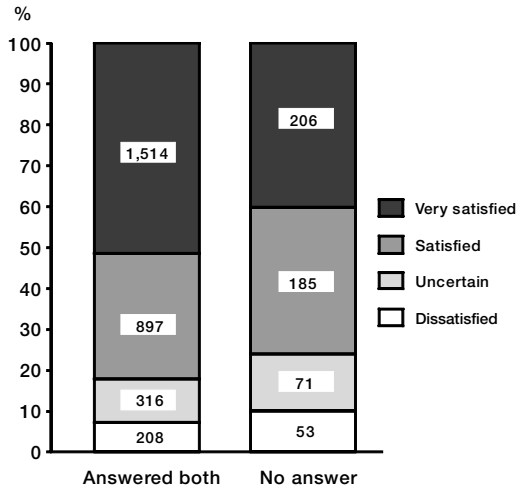

Figure 11. The relative percentage of satisfied patients in 1997. Patients that answered on both occasions (n=2935) were more satisfied than those that did not respond in 1998 (n=515) (Chi square p < 0.001).

the groups to be discriminated. For comparison the discriminative ability of the Oxford-12 was also tested. The Mann Whitney U-test was also used to compare differences in questionnaire results between these 2 groups. The construct validity of the satisfaction questionnaire was determined by correlating the answers to that of the other more extensive questionnaires using Spearman's non-parametric correlation coefficients. The weighted Kappa coefficient was calculated using the level of satisfaction on an ordinal scale (1–4). A Kappa coefficient of 0.4– 0.6 was considered fair, 0.6–0.8 good, and >0.8 excellent. The Kruskal Wallis test was used to compare differences in mean outcome scores with satisfaction as the grouping variable.

Results

The satisfaction questionnaire posed to all 27,114 living patients in 1997 was answered by 95% of the patients. When the 3600 patients were asked to answer the satisfaction questionnaire a 2nd time, in combination with the longer health outcome questionnaires in 1998, only 84% answered the satisfaction questionnaire. The response rates for the various health outcome questionnaires varied from 85% to 87%, diminishing to 57%–77% if only fully completed questionnaires were included. The weighted Kappa for the satisfaction questionnaire was 0.64, which can be interpreted as good agreement quality (reliability)

Of the 3,583 patients asked the short question both in 1997 and 1998, 73 patients did not answer on either occasion while 2,935 patients answered on both occasions. The 515 patients that answered the short question in 1997 but not in 1998 were older (Student-t, p<0.001, 95% CI 1.9–3.2 years), more often women (Chi-square, p=0.04) and more often unsatisfied in 1997 (MW, p<0.001) than those that answered on both occasions (Figure 11). There were an additional 60 patients that did not answer the first inquiry in 1997 but answered in 1998, but also among these there was a higher proportion of unsatisfied patients. The short question on satisfaction and the Oxford-12 questionnaire were found to have similar areas under the ROC Curve of 0.628 and 0.632, respectively (Figure 12). The revised patients were not as satisfied with their knee as those unrevised (MW, p<0.001) and their mean Oxford-12 score was worse (mean score=30/60) than that for the unrevised (mean score=25/60) (MW p<0.001). The satisfaction questionnaire had the highest correlation with the disease specific scores followed by those domains in the general health questionnaires that related to pain and to physical function (Table 5). For emotional parameters, the correlation was much lower.

Conclusions

The single-item satisfaction questionnaire has acceptable reliability, responsiveness and construct validity, hence meeting the basic requirements for psychometric validation. Furthermore, the response rate for the satisfaction questionnaire alone is higher than for the longer health outcome questionnaires, but this response rate decreases when it is coupled with the longer questionnaires.

When knee arthroplasty patients state, in a postal survey, that they are satisfied with their knee, they are mainly referring to the fact that they have gained good pain relief and improved function. When inquiring about the results of a treatment, in which the general benefit has already been proven and a preoperative health score is not known, a knee surgeon might be just as interested in patient satisfaction as in a score resulting from a more elaborate health outcome questionnaire.

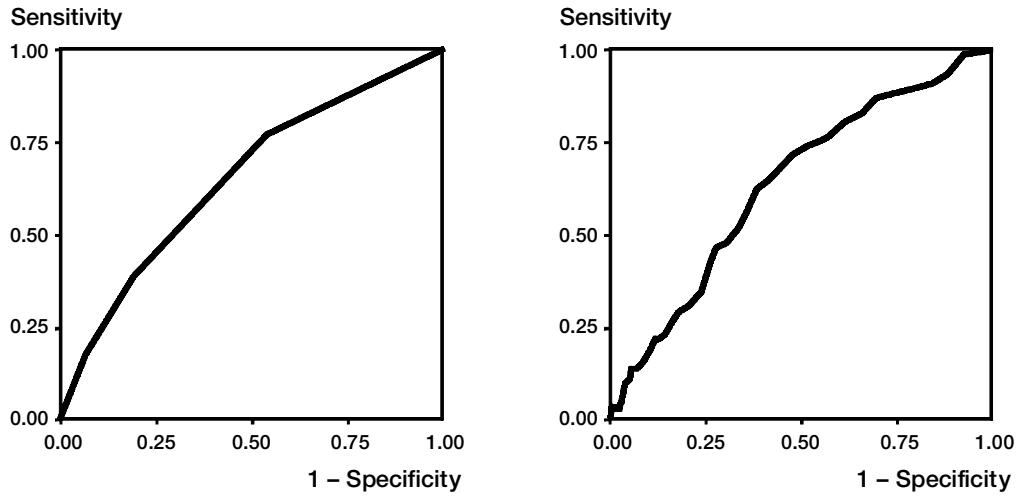

Figure 12. ROC curves indicating the ability to discriminate between revised and un-revised patients (n=823 and n=76). This is expressed as the area under the curve which is 0.628 for the single item satisfaction questionnaire (left) and 0.632 for the Oxford-12 (right).

Table 5. Correlation between patient satisfaction and different domains of general health and disease-specific questionnaires. P<0.001 for all correlation's

Questionnaire	Spearman	n
NHP		
Pain	0.62	669
Physical Mobility	0.47	690
Energy	0.42	711
Emotional Reaction	0.36	674
Sleep	0.33	702
Social Isolation	0.20	699
SF-12		
Physical Component Summary	0.42	579
Mental Component Summary	0.25	579
SF-36		
Body Pain	0.48	704
Physical Component Summary	0.45	485
Physical Functioning	0.43	628
General Health	0.39	666
Social Functioning	0.38	687
Vitality	0.35	684
Mental Health	0.34	686
Role-Emotion	0.32	687
Mental Component Summary	0.32	485
Role-Physical	0.29	693
Oxford-12	0.68	899
WOMAC		
Pain	0.67	957
Physical Function	0.64	854
Stiffness	0.63	977

Paper IV: What's all that noise? The effect of co-morbidity on health outcome questionnaire results after knee arthroplasty

Introduction

The Orthopedic community is increasingly relying on health outcome questionnaires to define and contrast the value of joint replacement surgery. However, questionnaires are imperfect and their results can be confounded by noise from sources other than the signal of interest. Sources of noise include age, gender, pre-operative diagnosis, and co-morbidity. Without recognizing and controlling for the sources of noise, the value of questionnaires for assessing outcomes after arthroplasty is suspect (Gross 1988).

Charnley recognized the importance of accounting for co-morbidity when assessing outcomes after hip arthroplasty and advocated stratifying patients by degree of co-morbidity to allow for meaningful comparisons. The resulting patient strata represent a functional classification and are often referred to as the "Charnley Class". Previously, results of health outcome questionnaires applied to hip arthroplasty patients were found to be significantly influenced by Charnley Class (Garellick et al. 1998).

The effect of Charnley Class, or co-morbidity, on the results of health outcome questionnaires

applied to knee arthroplasty patients has not been well defined. Therefore, the purpose of this study was to first modify the Charnley Classification for application to knee arthroplasty patients and then determine what effect co-morbidity, as defined by the modified Charnley Class, had on the results of a spectrum of outcome questionnaires. The hypothesis was that general health questionnaires would be influenced by modified Charnley Class, disease specific questionnaires less so, joint specific questionnaires minimally, and a single item questionnaire about the index knee not at all.

Methods

A postal survey was sent to 3600 patients randomly selected from the SKAR (Paper II). All patients were sent 1 of 4 general health questionnaires in combination with 1 of 3 disease/site specific questionnaires. All 3600 patients were also sent the single-item satisfaction questionnaire, a questionnaire regarding patient burden, and 2 single-item questionnaires regarding their index knee (Single-Item Knee Score) and the other regarding their overall health (Single-Item Health Score). The Single-Item Knee Score asked the patient to rate their impression of how their index knee felt on a scale of 1 to 10, and the Single-Item Health Score asked the patient to rate the impression of their general health on a scale of 1 to 10. For both questionnaires a score of 1 represented the best possible score and a score of 10 represented the worst. The modified Charnley Class questionnaire was also sent.

Multinomial regression was performed to determine the variables that affected modified Charnley Class. Gender was used in the regression as a factor with patient age at the time of postal survey and the year of operation as covariates. ANOVA was used to compare mean ages between Charnley Classes while the Chi Squared test was used to compare the frequency distribution of Charnley Class by gender and by age category (<75 years and ≥ 75 years). Differences in questionnaire scores by modified Charnley Class were determined with the Kruskal Wallis test. *P*-values of less than 0.05 were considered significant. Linear regression analyses were performed for each questionnaire with the questionnaire score as the dependent variable and patient age at the time of postal survey, gender, time since operation, type of prosthesis (uni-compartmental versus total), revision status, and modified Charnley Class as the independent variables. Logarithmic transformations were performed to normalize the distribution of skewed scores when performing linear regressions.

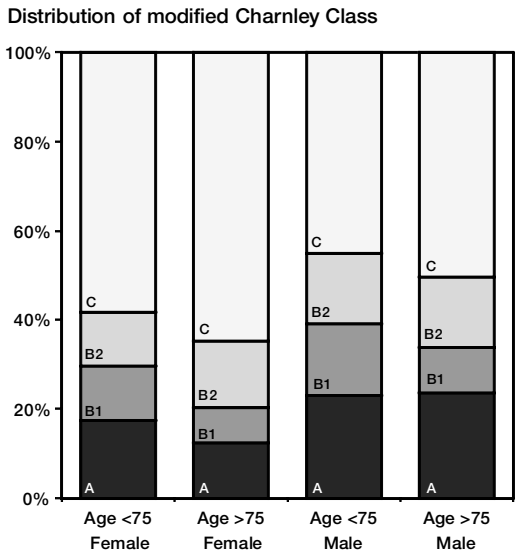

Figure 13. Distribution of modified Charnley Class by age and gender.

Results

Multinomial regression demonstrated that gender and patient age at the time of mail-out significantly affected the modified Charnley Class distribution ($p < 0.001$). ANOVA confirmed the differences in age between Charnley Classes, but the differences were clinically small (maximum difference 2 years) and were only significant for females.

The distribution of Charnley Classes differed between females and males ($p < 0.001$) with females having a higher proportion of patients in Charnley Class C even after age distribution had been accounted for. While there was no difference in the distribution of Charnley Classes between age groups for males, females younger than 75 years had a different distribution compared to those 75 years and older ($p<0.001$) with older females having a higher frequency of Charnley Class C patients (Figure 13).

For all questionnaires tested, significant differences were found in the scores when analyzed by

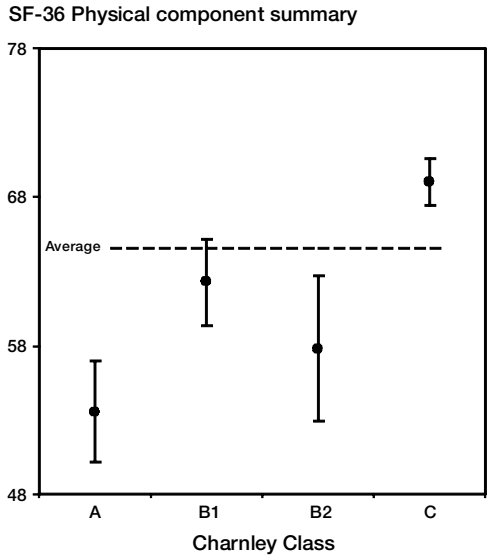

Figure 14. Variation in SF-36 Physical Component Summary scores by Charnley Class for females < age 75. Error bars represent 95% confidence intervals. Range of scores listed on the Y-axis represent 2 standard deviations. N.B. scores have been inverted for comparative purposes.

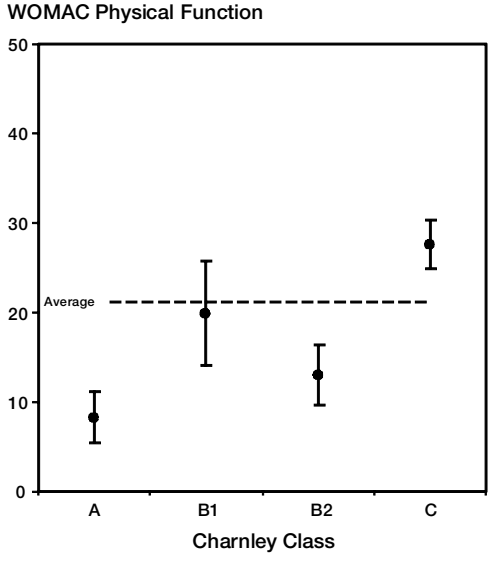

Figure 15. Variation in WOMAC Physical Function scores by Charnley Class for females < age 75. Error bars represent 95% confidence intervals. Range of scores listed on the Y-axis represent 2 standard deviations.

Charnley Class. A consistent pattern emerged for the distribution of scores by Charnley Class (Figures 14–16). Patients with mono-articular knee involvement, treated with arthroplasty (Class A)

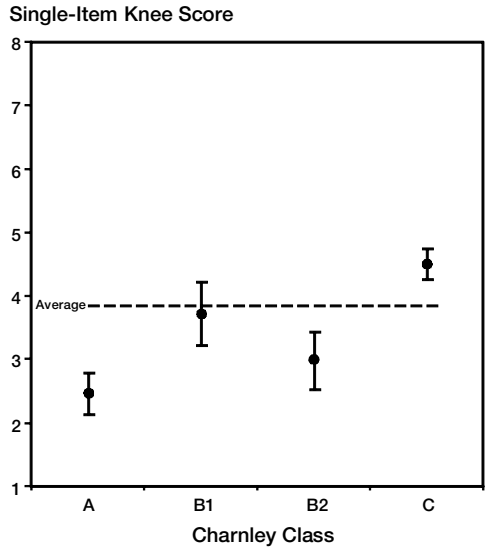

Figure 16. Variation in Single-Item Global Knee scores by Charnley Class for females < age 75. Error bars represent 95% confidence intervals. Range of scores listed on the Y-axis represent 2 standard deviations.

scored the best while patients with 1 arthroplasty and arthritis in the contralateral knee (Class B1) scored significantly worse. Patients with bilateral arthroplasties (Class B2) tended to score as if they had no arthritis in the knee contralateral to the index knee (i.e. Class A). Patients with knee arthroplasty and remote arthritis or systemic disease affecting their ability to ambulate (Class C) scored worse than all other classes. These results were found regardless of the type of questionnaire or stratification of scores by gender or patient age.

While a consistent pattern in questionnaire scores by Charnley Class was noted, the magnitude of the change varied by questionnaire (Table 6). The WOMAC scores varied the most, with a 75% increase in Physical Function scores when comparing Charnley Class A to B1, and a 138% increase from class A to C. The Oxford-12 scores varied to a lesser degree with a 34% increase in scores between Charnley Class A and B1 and a 55% increase between Charnley Class A and C. Similar changes were noted for the Single-Item Knee and Single-Item Health Scores. However, the Single-Item Knee and Single-Item Health scores had less of a change between Charnley Class A and B2. The SF-36 Physical and Mental

Table 6. Percentage change in questionnaire scores by Charnely Class (all patients)

Questionnaire	% change in Charnley Class		
	A to B1	A to B2	A to C
SF-36 Physical Comp. Sum. (n=484)	12.8	7.7	22.1
SF-36 Mental Comp. Sum.(n=484)	-1.6	-3.9	10.4
Sigle-Item Health Score (n=2736)	19.9	1.8	58.1
WOMAC Pain (n=934)	67.0	22.7	126.6
WOMAC Stiffness (n=951)	72.8	19.7	106.9
WOMAC Physical Function (n=836)	75.2	35.0	138.0
Oxford-12 Knee Score (n=882)	33.9	18.1	54.7
Single-Item Knee Score (n=2773)	29.2	5.0	54.5

Component Summary scores changed the least by Charnley Class.

Linear regression analyses for the various scores tested demonstrated a variety of covariates as having an effect on the scores, depending on the questionnaire (Table 7). However, for every questionnaire the modified Charnley Class was a significant factor, even when all other factors were accounted for in the regression equation ($p<0.001$). No other factors were significant for all questionnaires.

Conclusions

Co-morbidity had a significant effect on outcome questionnaires after knee arthroplasty, regardless of the specificity of the questionnaire used. Results of questionnaires varied by as much as 138% between Charnley Classes. Co-morbidity should be accounted for in outcome studies, especially with a discriminative questionnaire application. The modified Charnley Classification questionnaire for knee arthroplasty is a useful method for assessing co-morbidity in this population. In essence, it is not possible to isolate the knee with health outcome questionnaires.

Table 7. Results of linear regression demonstrating significant factors that effect scores of health outcome questionnaires applied to knee arthroplasty patients

Questionnaire	n	Transf.*	Factor	p value
SF-36 Physical Comp. Sum.	484	None	Charnley	<0.001
			Age at survey	<0.001
			Gender	0.013
			Type (Uni. Vs Total)	0.026
SF-36 Mental Comp. Sum.	484	None	Charnley	0.001
Single-Item Health Score	2736	None	Charnley	<0.001
			Age at survey	<0.001
			Gender	<0.001
			Operative year	0.008
			Revision status	0.048
WOMAC Pain	934	log10	Charnley	<0.001
			Revision status	<0.001
			Gender	0.004
WOMAC Stiffness	951	None	Charnley	<0.001
			Revision status	<0.001
WOMAC Physical Function	836	None	Charnley	<0.001
			Revision status	<0.001
			Age at survey	0.016
			Operative year	0.004
			Gender	0.011
			Type (Uni. Vs Total)	0.028
Oxford	882	log10	Charnley	<0.001
			Operative year	<0.001
			Revision status	0.024
			Type (Uni. Vs Total)	0.033
Single-Item Knee Score	2773	log10	Charnley	<0.001
			Revision status	<0.001

* Transformation required to normalize regression residuals plot.

Paper V: Translation and validation of the Oxford-12 Item Knee Score for use in Sweden

Introduction

The Oxford-12 Item Knee Score was a new and well-validated outcome questionnaire designed for use with knee arthroplasty patients. The Swedish translated version of the Oxford-12 performed optimally across multiple parameters in a cross-sectional study (Paper II). However, it is insufficient to solely translate a questionnaire into a foreign language without validating the translated version (Guillemin et al. 1993, Mathias et al. 1994). Therefore, the purpose of this study was to translate and validate the Oxford-12 for use in Sweden.

Methods

The Oxford-12 standard English version was independently translated into Swedish and back translated by a professional translator and a bilingual Orthopaedic surgeon. Adequacy of the translated versions was assessed and a final translated version was agreed upon. A pilot study was conducted on 8 bilingual subjects who completed in random order the Swedish and English version of the Oxford-12, separated by a 5-day interval, to further asses the translation.

A 1200 patient subset of the 3600 patients randomly selected from the SKAR (Paper II) was used. The subset represents all patients who received the NHP, SF-12, SF-36, or SIP in combination with the Oxford-12. Inclusion criteria are the same as for Paper II, above. A cover letter was included along with a postage-paid return envelope and a 3rd questionnaire regarding patient burden. A reminder letter was sent at 2 weeks for non-responders. At 3 weeks, 120 patients were randomly selected from those that completed the Oxford-12 and were sent a WOMAC.

Feasibility was determined by calculating the percentage of questionnaires returned and the percentage of questionnaires that were returned comprehensively completed. Missing responses were not imputed.

Convergent and divergent construct validity were tested by examining the Spearman's correlation coefficients of the Oxford-12 scores compared to the domains of the general health questionnaires and the WOMAC. It was hypothesized that the Oxford-12 should correlate highest with the physical and pain domains of the other questionnaires (convergent validity) and lowest with the Eating domain of the SIP and the psychosocial domains of the general health questionnaires (divergent validity). Content validity was investigated by examining the skew of the distribution as well as floor and ceiling effects.

To determine test-retest reliability, 60 patients were randomly selected from those who had completed the Oxford-12. Each was mailed a repeat Oxford-12 at 4 weeks. Both the ICC and the coefficient of repeatability were calculated (Bland et al. 1986).

Internal consistency was determined by calculating Cronbach's. A value for Cronbach's alpha greater than 0.8 was considered good while a value greater than 0.9 was considered excellent.

Discriminative ability was tested by comparing the Oxford-12 scores generated for revised and unrevised knees with the Mann Whitney U test and by calculating the area under the ROC Curve. The same tests were performed for the WOMAC. It was hypothesized that the WOMAC and Oxford-12 should have similar discriminative ability.

Results

The 2 translated versions of the Oxford-12 were very similar, and a common version was accepted incorporating aspects of both translations. Back translation of the accepted version was stable. The original and translated versions were judged to be culturally and linguistically equivalent.

On average, patients reported requiring 10 minutes to complete the questionnaire and 23% of patients stated that they required assistance to complete it.

Of the 1200 Oxford-12 questionnaires posted, 2 were returned by the post office for incorrect address and 3 were returned with a note by a family member or caregiver indicating that the patient was deceased. 1026 questionnaires were returned at least partially completed, yielding a response rate of 86%. Of these, 89% were complete. The net response rate therefore was 77%.

Table 8. Ability of Oxford-12 and WOMAC to distinguish between revised and unrevised knee arthroplasty patients

Questionnaire	n	Mann-Whitney U-test	Area under ROC [a] curve	95% CI for ROC curve	Asymptotic sig. ROC curve
Oxford-12	917	p < 0.0001	0.64	(0.58–0.70)	p < 0.001
WOMAC pain	967	p < 0.0001	0.70	(0.64–0.76)	p < 0.001
WOMAC stiffness	986	p < 0.0001	0.66	(0.60–0.72)	p < 0.001
WOMAC physical function	862	p < 0.0001	0.67	(0.60–0.74)	p < 0.001

[a] Receiver operating characteristic curve

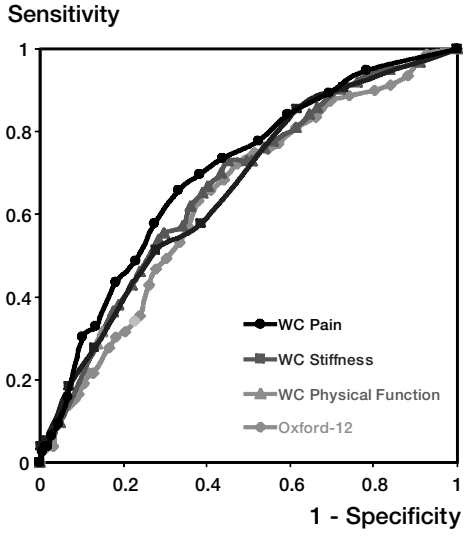

Figure 17. Receiver Operating Characteristic Curve demonstrating comparable ability of the Oxford-12 and WOMAC to discriminate between patients with unrevised and revised knee arthroplasties. Discriminative ability is related to the area under the curve.

The Oxford-12 correlated closely with the physical domains and less so with the mental and social domains in all general health questionnaires. Correlations with the WOMAC domains were the highest (Pain, Rho = 0.87, Stiffness, Rho = 0.83 and Physical Function, Rho = 0.74). The Oxford-12 correlated poorly with the Eating Domain of the SIP (Rho = 0.14) hence demonstrating good divergent construct validity.

6.8% of patients surveyed who completed the questionnaire recorded the best possible score. Only 0.1% recorded the worst possible score. The frequency distribution of the score was skewed to the right (better) with a skew value of 0.73.

The ICC for the Oxford-12 was high at 0.94 (95% confidence interval 0.89–0.96). The mean difference between the 2 sets of scores was –0.7 (95% CI –2.0–0.6), which was not significantly different from 0 (one sample t-test). The coefficient of repeatability was 9.6 and 95% of the values were within –0.7 ± 9.6.

The internal consistency was excellent with a Cronbach's alpha of 0.93 (95% confidence interval 0.63–0.84). Removal of any of the 12 items in the calculation of Cronbach's alpha did not result in a value greater than 0.93.

All 3 domains of the WOMAC discerned a difference between the unrevised and revised groups both with the Mann Whitney U test and the area under ROC Curve (Table 8, Figure 17). The Oxford-12 displayed similar ability using the same methods.

Conclusions

The Swedish translation of the Oxford-12 Knee Score is linguistically and culturally equivalent to the English version and it has solid psychometric characteristics in keeping with the original questionnaire. This translated version is appropriate for general use with knee arthroplasty patients in Sweden.

Paper VI: Post-operative patient disposition after knee arthroplasty based on preoperative WOMAC scores

Introduction

Health outcome questionnaires applied to the SKAR, to date, have been used in a cross-section-

al, discriminative fashion. Questionnaires have not been applied pre-operatively. The longitudinal nature of the questionnaires tested has not, therefore, been directly investigated.

In Canada and Sweden, waiting times for surgery are increasing as surgeons are forced to rationalize the delivery of knee arthroplasty. Furthermore, demand for knee arthroplasty is predicted to increase over the next three decades (Robertsson et al. 2000). However, there is no consensus regarding the prioritization of patients on a knee arthroplasty waiting list and furthermore, the effect of delaying the delivery of the surgery are unknown.

The Western Ontario and MacMaster Universities Osteoarthritis Index (WOMAC) is a well-validated and widely used health outcome questionnaire that has relevance for a knee arthroplasty population. The first purpose of this study was to determine the pre-operative WOMAC scores for patients on an elective total knee arthroplasty wait list and to determine the post-operative disposition of those patients based on their pre-operative WOMAC scores. The hypothesis was that patients scoring substantially worse on pre-operative WOMAC scores would not obtain the same post-operative WOMAC scores as the other patients. The second purpose of this study was to investigate which questions, if any, within the WOMAC accounted for the variation in the pre and post-operative scores.

Methods

156 primary total knee arthroplasties with a diagnosis of osteoarthritis were followed prospectively in a multicentre trial. The standard North American Version of the WOMAC was employed in a patient self-completed format preoperatively and 1 year post-operatively. Patients were prompted to complete missing responses and any residual deficiencies in responses were imputed. Only the Pain and Physical Function domains were used for the purposes of this study because of the significant floor and ceiling effect seen with the Stiffness domain.

Preoperative WOMAC scores were categorized into two ordinal groups for each domain. Group "Better" was defined as patients scoring one standard deviation below the mean score, and group "Worse" as patients scoring one standard deviation above the mean score. The two pre-operative groups were used as a factor in determining differences in post-operative WOMAC results using the non-parametric Mann Whitney test. The effect of age, gender, body mass index, patellar resurfacing status and co-morbidity on pre and post-operative WOMAC scores was determined using multiple regression.

To investigate the effect of pre-operative WOMAC scores on the change in pre and post-operative WOMAC scores, the scores were again grouped into the same two groups as defined above. A delta score was calculated for groups "Better" and "Worse" for each question within the Pain and Physical Function domain. Differences in delta scores for each question by group and domain were checked using the Kruskal Wallis test.

Results

Of the covariates tested, only gender had a significant effect on preoperative WOMAC scores ($p<0.05$). Patients scoring 1 standard deviation higher (group "Worse") on preoperative WOMAC scores for Pain and Physical Function had significantly worse post-operative scores for the respective domains (Pain $p=0.011$, Physical Function $p=0.023$, Figures 18 and 19). Despite the fact that the postoperative scores were significantly different, patients in groups "Worse" and "Better" had a similar net change for the Pain domain (60.0% change for group "Better" and 69.5% change for group "Worse"). The net change for the Physical Function domain differed by group with group "Worse" having a higher average change from pre- to postoperative (35.4% change for group "Better" and 60.6% change for group "Worse").

9.5% of patients in group "Better" were worse postoperatively compared to 0% in group "Worse" for the Pain domain (Figure 20). These differences did not reach statistical significance (Fisher exact test $p=0.08$). For the Physical Function domain, 37.5% of patients in group "Better" were worse postoperatively compared to 0% in group "Worse". These differences were significant (Fisher exact test $p=0.006$).

There was no difference in single item scores for the Pain and Physical Function domains for patients with group "Worse" WOMAC scores

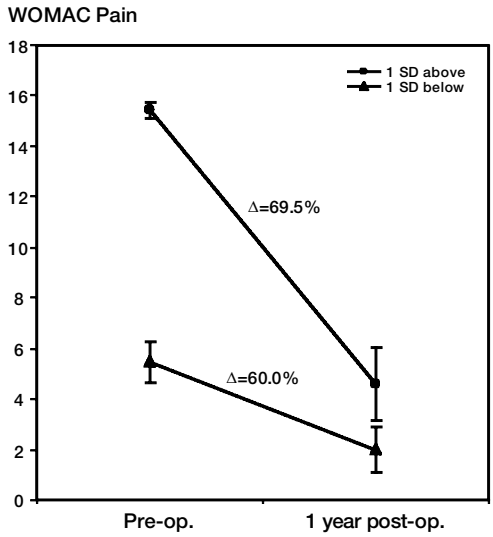

Figure 18. Post-operative disposition of WOMAC Pain scores when stratified by pre-operative score. Post-operative scores are significantly different (Mann–Whitney, P=0.11).

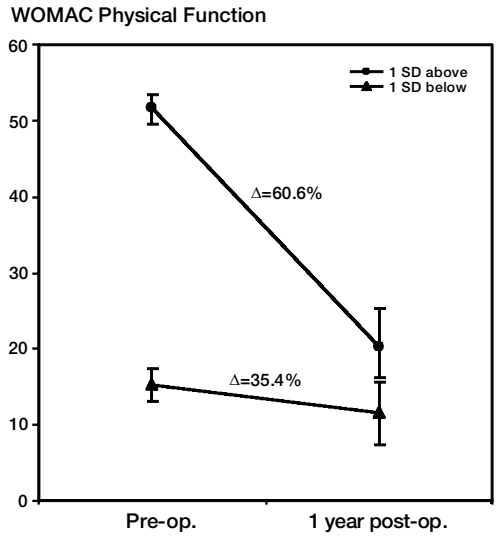

Figure 19. Post-operative disposition of WOMAC Physical Function scores when stratified by pre-operative score. Post-operative scores are significantly different (Mann–Whitney, P=0.23).

(Kruskal Wallis, p>0.05). However, for the same domains with group "Better" patients, there were significant differences between changes in single item scores for both the Pain (p=0.009) and the Physical Function (p=0.04) domains. Questions 1 and 2 (pain while walking on flat surface and pain going up or down stairs, respectively) for the Pain domain demonstrated the greatest change in scores while questions 3 and 4 (pain at night while in bed and pain while sitting or lying, respectively) demonstrated the least change (Table 9). Questions 1 and 2 for the Physical Function domain (descending stairs and ascending stairs, respectively) demonstrated the greatest change while questions 13, 14 and 15 (getting in/out of a bath, sitting, and getting on/off a toilet, respectively) demonstrated the least change.

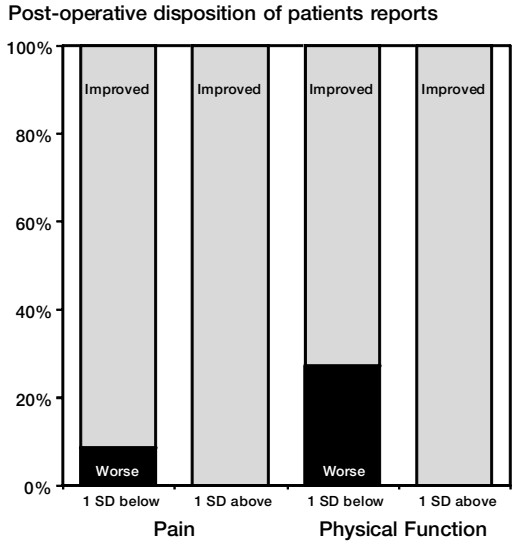

Figure 20. Post-operative disposition (Improved versus Worse) of patients reports of Pain and Physical Function for the WOMAC based on pre-operative score stratification.

Conclusions

When using the WOMAC to compare a relatively large group of knee arthroplasty patients, patients scoring significantly higher (worse) pre-operatively can not be expected to obtain the same absolute result, as measured by the WOMAC. However, some patients scoring significantly lower (better) on the WOMAC pre-operatively actually registered worse WOMAC pain and Physical Function scores post-operatively. The questions in the WOMAC Pain and Physical Function domains regarding ascending and descending stairs consistently registered the best net improvement in scores, regardless of pre-operative status.

Table 9. Change in WOMAC Pain and Physical Function Domain scores between pre and post–operative application broken down by question. Scores are stratified for each domain by the pre–operative score's relationship to the mean

	Group "Worse" 1 SD above the mean domain score			**Group "Better"** 1 SD below the mean domain score		
Domain and Question #	Δ pre-op. to post-op.	95% CI	% of change in domain score by question	Δ pre-op. to post-op.	95% CI	% of change in domain score by question
Pain						
Question 1	1.76	1.23–2.29	17.4	0.85	0.52–1.18	29.0
Question 2	1.92	1.49–2.35	19.0	1.04	0.61–1.47	35.5
Question 3	2.15	1.80–2.50	21.3	0.31	0.04–0.58	10.6
Question 4	2.19	1.86–2.52	21.7	0.27	0.02–0.52	9.2
Question 5	2.08	1.75–2.41	20.6	0.46	0.03–0.89	15.7
Physical Function						
Question 1	1.96	1.57–2.35	6.1	1.00	0.53–1.47	13.2
Question 2	2.15	1.78–2.52	6.7	0.85	0.40–1.30	11.3
Question 3	2.04	1.67–2.41	6.4	0.54	1.07–1.01	7.2
Question 4	1.96	1.51–2.41	6.1	0.42	0.09–0.75	5.6
Question 5	1.58	1.13–2.03	4.9	0.50	–0.05–1.05	6.6
Question 6	2.04	1.69–2.39	6.4	0.65	0.24–1.06	8.6
Question 7	1.77	1.42–2.12	5.5	0.62	0.27–0.97	8.2
Question 8	1.85	1.46–2.24	5.8	0.62	0.27–0.97	8.2
Question 9	2.04	1.65–2.43	6.4	0.19	–0.18–0.56	2.5
Question 10	2.00	1.63–2.37	6.2	0.46	0.07–0.85	6.1
Question 11	2.12	1.75–2.49	6.6	0.19	–0.22–0.6	2.5
Question 12	1.62	1.27–1.97	5.0	0.23	–0.14–0.6	3.0
Question 13	1.62	1.05–2.19	5.0	0.08	–0.29–0.45	1.1
Question 14	1.65	1.32–1.98	5.1	0.12	–0.21–0.45	1.6
Question 15	2.31	1.96–2.66	7.2	0.12	–0.27–0.51	1.6
Question 16	1.50	1.15–1.85	4.7	0.58	0.09–1.07	7.7
Question 17	1.88	1.61–2.15	5.9	0.38	0.03–0.73	5.0

Discussion

Paper I

The reasons for performing a knee arthroplasty and the goals to be achieved by the surgery are various. Typical objectives are to reduce pain and deformity as well as improve mobility and walking ability. Depending on the pre-operative status of the patient, a varying change in these factors can be expected. However, the ultimate goal for a treatment modality must be to have satisfied patients who remain so over the long-term.

By quantitating the subjective outcome of satisfaction, the validity of the satisfaction questionnaire as a measure of the condition of interest is questioned. Even though satisfaction is a commonly used concept, it is not a concept that can be directly measured, or can be validated against a specific criterion. Instead, the construct validity of the satisfaction score has to be tested by correlating satisfaction to the results of other defined measures such as more extensive health or disease/site specific questionnaires. It has been demonstrated previously, for example, that patient satisfaction after arthroplasty has a significant correlation to pain and to physical function (Anderson et al. 1996, Heck et al. 1998).

Although satisfaction may be affected by factors that seem unrelated to the surgical intervention (e.g. patient-surgeon relationship, attitude of staff, availability of parking spots, etc.), it can be assumed that patients' answers regarding their satisfaction with a treatment is of general interest to surgeons and that the questionnaire thus is warranted.

In some previous studies where patient satisfaction has been accessed, the percentage of satisfied patients has been quoted as 85-89% (Anderson et al. 1996, Hawker et al. 1998, Heck et al. 1998). In this study, the overall percentage of satisfied patients was 81%, but only 8% were unsatisfied while 11% remained uncertain, however, this study included a wider range of diagnoses and implants, which may account for the difference.

The pathology leading to arthroplasty significantly affected the level of satisfaction, with patients suffering long-standing disease being more satisfied. Assuming that the patients remember their own pre-operative status as a comparison when answering a question regarding satisfaction, this seems logical. A patient with chronic rheumatoid arthritis, for example, usually affecting several joints, has a different pre-operative function than a patient with osteonecrosis who probably experienced a sudden onset of pain and dysfunction in an isolated joint. Previously, it has also been shown that the absence of problems in the contralateral knee is a predictor of better physical function (Hawker et al. 1998). These findings illustrate the importance of taking the pre-operative condition of patients into account when evaluating clinical results.

The consistency regarding satisfaction in the unrevised cases over the 15 years shows that a successful knee arthroplasty can be expected to give a lasting good clinical result.

Patient satisfaction after TKA and UKA was similar. In the case of a revision, revised UKA's were more satisfied than revised TKA's. This can be partly explained by the fact that TKA is more prone to infections and related complications (Robertsson et al. 1999a). However, the advantage of the UKA is counteracted by the fact that the risk of revision is lower for the TKA.

Patients with patellar resurfacing were found to be more satisfied than patients without. The use of patellar components in TKA has long been a matter of debate. Some authors claim an advantage of a patellar resurfacing (Schroeder-Boersch et al. 1998), while others fail to find such advantage (Barrack et al. 1997). The cause of the different findings might be explained by the finding that the benefit of the patellar component diminishes with time.

Not surprisingly, it was found that revised patients were less satisfied than those unrevised. One would expect that being subjected to two or more operations affected the level of satisfaction.

That only 22% of cases were dissatisfied with their knee after revision must be considered as an indicator of the benefit of the revision surgery.

Paper II

This study has avoided comparing the construct validity of the questionnaires tested because of the potential for circuitous and sophistic logical traps. Therefore, it was decided to concentrate on the content validity of each questionnaire. Comparing the responsiveness of the questionnaires tested was also avoided, as the purpose of this study was to define questionnaires that would be appropriate for a cross-sectional and discriminative postal survey. This study was intentionally not limited to tricompartmental or primary arthroplasties so that the results would be applicable for a wide range of patients registered with the SKAR.

Previous comparative studies have been published investigating various aspects of specific outcome questionnaires. All questionnaires tested had higher than expected response rates compared to other published results (Asch et al. 1994, McHorney et al. 1994a, Plant et al. 1996). Stucki et al. (1995) compared the SF-36 to the SIP on 54 patients undergoing elective total hip replacement. They also found large floor effects for the SIP and concluded that it was a less relevant questionnaire than the SF-36 for total hip arthroplasty. This agrees with our results. Beaton et al. (1997) investigated the reliability and responsiveness of five general health questionnaires as applied to workers with musculoskeletal complaints. The questionnaires tested included the NHP, SF-36 and SIP. Reliability estimates (ICC's) for these questionnaires were slightly higher than the findings reported here; perhaps reflective of the younger patient population studied. However, their reliability estimates ranked in the same order as the results reported here (NHP > SIP > SF-36). Essink-Bot et al. (1997) compared four general health questionnaires, including the SF-36 and NHP, on a population suffering from migraines. They found the NHP to have better feasibility, but a more skewed distribution with a larger percent of minimum scores and lower internal consistency (Cronbach's Alpha) than the SF-36. These findings are in complete agreement with the findings from this study.

A methodological approach was taken in this study in order to rationalize the choice of appropriate questionnaires for future application to the SKAR. The questionnaires deemed most appropriate, the SF-12 and Oxford-12, were so only when factoring with equal weight all the criteria tested, including feasibility, burden, content validity, and reliability. It is likely that these questionnaires may not be the most appropriate for other types of applications, such as an evaluative postal survey, or when the different parameters are weighted differently. A methodological approach yielded useful data and is worthwhile when investigating candidate questionnaires.

Paper III

It is important for every surgeon to have some information regarding the results of their interventions. However, using extensively tested and validated health outcomes questionnaires to inquire about post-operative status is not without difficulties. To be meaningful, a score produced by a questionnaire has to be compared to some kind of metric, such as the pre-operative score, or the score of a matched otherwise healthy cohort. Unfortunately, standardized questionnaire results for comparable cohorts are not widely available, particularly for elderly knee arthroplasty cohorts. Additionally, it cannot be automatically assumed that the operation was meant to restitute the knee to that of a completely "healthy" individual, not to mention the general health. Thus, if a pre-operative score is not known, it is difficult to decide from a post-operative score alone what the strengths and weaknesses of the intervention were. This is in line with the findings of Brinker et al. (1997) who concluded that observed differences in knee scores between study groups were at least as likely to represent differences in the patient populations as the differences in the operative technique or design of the implant.

In longitudinal studies a directional change in an outcome score has a meaning, while in cross-sectional studies the raw numerical value of a score can be difficult to interpret in isolation. This

has led authors to convert the raw numeric score into nominal categories (Insall et al. 1976). A certain range of scores thus becomes classified as excellent, another good, etc. Such arbitrary categorization is often post hoc and although it may be valid for a specific population, it can not be generalized. Such generalization also leads to a reduction in statistical power.

In lieu of a standardized metric of operative success, Orthopaedic surgeons performing knee arthroplasties have often asked their patients if they are satisfied with the operated knee. Patient satisfaction is admittedly a subjective description that is based on a variety of factors. However, the assumption can be made that when asked about post-operative satisfaction, patients relate their perceived surgical result to that expected of the operation, even though the knee function is not necessarily comparable to that of a healthy subject.

In the postal surveys, it was found that a single-item satisfaction questionnaire had a high response rate and good reliability. Furthermore, more patients answered the short question than the more extensive questionnaires, and those that did not respond were not a random subset of the population regarding satisfaction, age or sex. The short question on satisfaction was as good as the validated Oxford-12 knee score in discriminating between previously revised and unrevised patients. Correlation between satisfaction and both general-health and disease specific scores was found, which is in agreement with Anderson et al. (1996) findings for the WOMAC and SF-36. However, the strength of the correlation varied, with the highest correlations seen for pain related domains followed by physical domains. Patients with the same level of satisfaction represented a wide range of results for health questionnaires, which indicates difficulties when interpreting health scores.

Having been subject to intensive testing, regarding properties such as reliability, responsiveness and validity, many extensive outcome questionnaires seem feasible for a variety of uses. However, it was demonstrated that psychometric validation methodology can also be applied to a simple questionnaire on satisfaction, which illustrates that successful testing and validation of a measure is not a panacea for easy interpretation or usefulness of results.

It should be stressed that the intention of this paper was never to advocate the replacement of well-known and respected health questionnaires with the single-item satisfaction questionnaire. The intention was only to evaluate the relation of patient satisfaction to some known validated measures. However, in the process of doing this, some interesting questions were raised. What is a reasonable measure of achievement after knee arthroplasty? Is a measure of the general benefit of surgery also a good measure when comparing different types of surgery, implants, etc.?

Although answers to these questions are not obvious, it can be concluded that when patients who have undergone knee arthroplasty, in a postal survey state that they are satisfied with their knee, they were mainly referring to the fact that they have gained good pain relief and improved function. Furthermore, when inquiring about the results of a treatment, in which the general benefit has already been proven and a pre-operative health score is not known, a knee surgeon might be just as interested in patient satisfaction as in the results of a more elaborate health questionnaire. The subtleties of prosthetic intra-design differences are lost in the positive effects of the knee arthroplasty intervention, as such.

Paper IV

Patient co-morbidity, as stratified by the modified Charnley Classification, was a significant factor for all questionnaires tested, regardless of the specificity of the questions to the index knee. This was an unexpected finding. In order to be certain that these result were not a function of different age or sex distributions for modified Charnley Class, data were analyzed while stratifying by these variables for the Kruskal Wallis test, and by including them along with other covariates in the regression equation. After accounting for all foreseeable sources of error, it was still found that Charnley Classes significantly affected the results of questionnaires.

Statistically significant changes in questionnaire scores by Charnley Class do not necessarily

imply clinically significant changes. To assess the quantitative impact, the percentage change in scores by Charnley Class was investigated. WOMAC scores more than doubled by Charnley Class, while the Oxford-12 and Single-Item Knee and Health scores increased by as much as 55%. Clearly, these changes would be clinically relevant. The SF-36 Physical and Mental Component Summary scores varied to a lesser degree. It is unclear if changes in these scores would be clinically relevant.

It could be assumed that the general questions within the SF-36 regarding concepts such as body pain and physical function would be susceptible to the "noise" of co-morbidity when inquiring about the index knee. Hence, the significant differences between Charnley Classes for the SF-36 Physical Component Summary were predictable. The fact that the changes in score by Charnley Class were small and questionably clinically relevant probably refers to the fact that there are no specific questions regarding the knee in the SF-36. Therefore, the signal for knee pathology in this questionnaire can be assumed to be low to begin with.

The disease specific WOMAC questionnaire inquires about pain with activity and the ability to perform activities such as stair climbing, putting on shoes and socks, etc. The noise of remote arthritis could be expected to impact on the WOMAC scores, as hip or spine arthritis could cause referred pain and interfere with a patient's ability to complete these tasks. The Oxford-12 score asks more specific questions related to the knee. In this case, less variation in scores by modified Charnley Class could be expected. This could account for the difference in the magnitude of the change in scores. Still, the Oxford-12 score was susceptible to the noise of co-morbidity. However, closer inspection of the Oxford-12 reveals that it too asks questions concerning stair climbing and putting on shoes and socks, hence it too can be rationalized to be susceptible to noise.

In an effort to concentrate singularly on the index knee, and to remove any extraneous questions that may pick up on remote arthritis or systemic disease, all patients were asked a single question regarding how their index knee felt on a scale of 1-10. Surprisingly, the same pattern occurred as for the other questionnaires and it was again found that there were significant differences in this score when compared by Charnley Class. Furthermore, the same magnitude of change in score occurred with this questionnaire as seen with the Oxford-12.

The effect of co-morbidity on surgeon-derived scores for knee arthroplasty patients (e.g., Knee Society Score, Hospital for Special Surgery Knee Score) have been previously investigated (Brinker et al. 1997). Patients having two or more significant medical conditions were found to have worse scores than others without the same level of co-morbidity. Furthermore, the authors concluded that when analyzing groups, without matching for sources of noise, differences in common knee scores between the groups are at least as likely to represent differences in the patient populations as in their treatments (Brinker et al. 1997). This is in general agreement with our findings, although our study shows that both remote arthritis and medical conditions affect patient derived outcome scores.

Garellick et al. (1998) found that the Charnley Class for hips significantly influenced the results of outcome scores applied to hip arthroplasty patients. This too is in agreement with our results for knee arthroplasty patients. Dawson and co-workers investigated the effect of remote joint co-morbidity on the change in the SF-36, Arthritis Impact Measurement Scales and the Oxford-12 Item Hip Score from pre and post-operative application (Dawson et al. 1996b, Dawson et al. 1996c). They found that the Oxford-12 Item Hip Score did not detect any difference between groups with and without remote arthritis, while the other questionnaires did. Based on this, they concluded that the Oxford-12 Item Hip Score was highly joint-specific and was not susceptible to the noise of remote arthritis. However, it should be emphasized that the differences between these patient groups generated by the remote arthritis (noise) may have been lost in the profound change in scores seen between pre-operative and post-operative patients (signal), regardless of the co-morbid status (Laupacis et al. 1993, Dawson et al. 1996b). This could explain the discrepancy between results in this paper and theirs, especially since the Oxford-12 Item Knee Score was applied in a discriminative fashion in this study.

The implication of this study is that the mind and body are one. Subsequently, it is not possible to assess the knee joint with questionnaires in isolation from the rest of the body, but instead, co-morbidity must be accounted for. This is particularly true when patients are evaluated in a discriminative fashion. Without such knowledge, erroneous conclusions could be drawn because of the significant impact that co-morbidity has on questionnaire results. The Charnley Class questionnaire that was employed seems like a convenient and effective way to assess patient co-morbidity when applying outcome questionnaires to knee arthroplasty patients.

Paper V

It is insufficient to simply translate a questionnaire into another language (Guillemin et al. 1993, Guyatt 1993). Instead, a more extensive approach is required in which cultural and language equivalence, as well as psychometric soundness, are checked. The Oxford-12 is a relatively concrete questionnaire, hence, cultural and language equivalence were anticipated and subsequently found to be maximal.

Patient burden imposed by administering the Oxford-12 was minimal, while the feasibility properties were maximal.

The Swedish translation of the Oxford-12 has been shown to be psychometrically sound. As expected, good convergent and divergent construct validity was demonstrated by the Spearman's correlations to the other questionnaires tested. These correlations mirror those reported by Dawson et al. (1998) for the English validation of the Oxford-12.

The translated version of the Oxford-12 had a definite floor effect but little ceiling effect and a moderate skew to the right (i.e., most patients reported good results). This is reflective of the overall favourable post-operative status afforded to the patients by the arthroplasty intervention. The floor effect and skew were, however, acceptable (Brazier et al. 1992, McHorney et al. 1994b, Martin et al. 1997). Still, logarithmic transformation of the scores should be considered when performing statistical tests (Bland 1995).

Both the ICC and the coefficient of repeatability (Bland et al. 1986) showed good test-retest reliability. The coefficient of repeatability was higher than that published by Dawson et al. (1998), but this may reflect the larger sample size and higher average patient age in this study. The internal consistency of the translated version of the Oxford-12 was excellent (Feinstein 1987). Identical values as reported by Dawson et al. (1998) for their post-operative patients were found.

Because of the cross-sectional nature of this study, classic measures of responsiveness were not applicable (Hays et al. 1993). The ROC Curve method had instead been used as an indirect measure of responsiveness (Essink-Bot et al. 1997). The WOMAC and Oxford-12 have comparable discriminative ability. Since the WOMAC has been previously found to be responsive using more conventional metrics (Roos et al. 1998), then these similarities suggest that the Oxford-12 would be equally responsive.

Dawson et al. (1998) were able to directly compute responsiveness using the effect-size (Kazis et al. 1989) with pre-operative and post-operative Oxford-12 scores. An effect size of > 0.8 is considered large, and Dawson et al. reported a profound effect size of 2.0. Because of the psychometric similarities between the English and Swedish Oxford-12 Knee Scores, an effect size greater than 0.8 between pre and post-operative applications of the Swedish Oxford-12 is likely. Therefore, the lack of a direct responsiveness statistic should not preclude the general use of the Oxford-12 in Sweden at this time. Validity is usually a matter of degree rather than an all-or-none property, and validation is an unending process (Nunnally et al. 1994).

Paper VI

Patients who score one standard deviation above (worse than) the mean pre-operative score have a less favourable impression of the health status of their knee post-operatively than patients scoring one standard deviation below (better than) the mean do. These patients do not obtain the same absolute post-operative Pain and Physical Function WOMAC scores and continue to have a less

favourable impression of the health status of their knee. Nevertheless, these patients have generally the same net improvement in their perception of Pain and nearly double the improvement in their perception of Physical Function.

Assuming that the natural history of primary knee osteoarthritis is for continued deterioration of the joint with a concomitant worsening in WOMAC scores, the use of the WOMAC as a prioritization tool for elective knee arthroplasty waitlist management seems justified. This is supported by the fact that all the patients in group "Worse" for Pain and Physical Function scores had improvement in the respective domain, while some patients in group "Better" scored worse post-operatively for Pain, and a significant number scored worse for Physical Function.

Still, while the use of the WOMAC in this capacity is appealing, the limitations stem from the lack of a logical cut point for prioritizing one patient over another and more importantly, from the fact that the comparisons in this paper are on a group-to-group basis, not on individual patients. Whether or not the psychometric properties of the WOMAC when used for this proposed application remain valid need to be tested in further studies.

Certain questions within the Pain and Physical Function domains of the WOMAC appear to represent a threshold for which knee arthroplasty is performed. For both the Pain and Physical Function domains, these questions relate primarily to ascending and descending stairs. While patients in group "Worse" had marked improvement across all questions within the Pain and Physical Function domains, it was only the questions relating to stairs and ambulation that demonstrated marked improvement for patients in group "Better". Patients in group "Better" had little or no improvement regarding pain at night or pain while sitting, and little or no improvement regarding the ability to bath, toilet, or sit. Future consideration for item reduction of the WOMAC when applied to knee arthroplasty patients should take these findings into account. Similar findings for the hip have been reported by Söderman et al. (2000).

General discussion

The lack of a gold standard for knee arthroplasty

A fundamental challenge when assessing outcomes after knee arthroplasty is the lack of a criterion, or gold standard, by which to compare and contrast the metric of interest. For example, what should the average patient score on the Oxford-12 with a TKA at 5 years versus 10 years be? Should the score decrease with time in an otherwise well functioning knee? If so, how would the change in the patient's age and overall physical condition affect the score? Should a UKA score be better or worse than a TKA? The answers to these questions are not at all obvious from the literature. Subsequently, researchers must compare the metric of interest against a hypothetical construct. For example, a patient who scores poorly on the WOMAC should score poorly on the Oxford-12. However, if the WOMAC has itself been validated against another construct, such as the Body Pain and Physical Function domains of the SF-36, the epistemological conundrum of outcomes research becomes apparent. Unlike Descartes Meditations, there is no "*cogito ergo sum*" or indisputable ground on which to base outcomes research. What precisely then is being measured when health outcome questionnaires are applied to knee arthroplasty? The short answer to this quandary is that nothing is being measured "precisely" with questionnaires. Instead, the questionnaires represent an imperfect attempt to quantify a largely qualitative phenomenon. This on the surface is somewhat discouraging. Still, attempting to quantify a patient's condition with a questionnaire improves the researchers understanding from that of a "meager and unsatisfactory kind" (Thompson 1910).

Most medical researches have a background in the sciences and at the very least are familiar with precise and reliable metrics for items such as hemoglobin, pulmonary artery pressure and weight, for example. Subsequently, most would be more comfortable working with criterion based metrics as opposed to construct based metrics. The field of health outcomes research generally, and specifically for arthroplasty, is strongly dependent on construct based metrics. Embracing health out-

comes research therefore results in a departure from the firm footing of the criterion to the uncertain ground of the construct. This can be initially quite disconcerting.

Why does knee arthroplasty, or surgery in general, lack a gold standard? The answer lies in the reflection that questionnaires are applied to the person, not the cell nor the organ, nor the joint. When researchers design and apply a questionnaire on pain after knee arthroplasty, for example, an imperfect metric is applied to an uncertain clinical picture that is highly influenced by all manner of psychosocial interactions occurring within the subject. The questionnaires may be picking up on the patient's satisfaction regarding how close they could park to the clinic door, a recent death in the family, or perhaps, the placebo effect imparted by the surgeon and the procedure. This apparent deluge of noise is not, however, necessarily a negative event. Instead, the supposed noise may in fact represent a portion of the signal of interest, the signal of the art, or humanistic side, of healing. This is partially what the health outcomes researcher is interested in. Therefore, the thoughtful researcher should be aware of the limitations of outcomes questionnaires and do everything possible to amplify the signal of interest while at the same time reducing the noise in the metric.

Discriminative versus evaluative outcomes studies

Conceptually, outcome questionnaires can be applied in 3 ways: predictive, discriminative, and evaluative (Kirshner et al. 1985). A predictive application is useful when a gold standard is known and the questionnaire, in effect, functions as a diagnostic or screening tool. A predictive application in not applicable to this thesis as there is no gold standard for knee arthroplasty. A discriminative application is used to differentiate between groups, while an evaluative application is used to measure the magnitude of a longitudinal change in the condition of interest. The two latter applications do not rely on a gold standard; however, the evaluative application relies on longitudinal data, which so far is not available with the SKAR. Therefore, the only relevant application to this thesis is for a discriminative application. In order for a questionnaire to be robust for application in a discriminative fashion, the questionnaire should demonstrate certain properties, which may not be complementary for a longitudinal application.

A dichotomous item scale is more appropriate for a discriminative application, while a polychotomous item scale is better for an evaluative application. For example, if a questionnaire aims to discriminate between a revised and unrevised knee, and if the items within asks, "do you have pain in your knee when climbing stairs?", a "yes" or "no" response scale forces the respondent to choose one answer. Either they have pain or they do not, and the resulting answer is clear. However, in the polychotomous item scale, such as "no pain, mild pain, moderate pain, and severe pain", the resulting answer is subject to patient variability in the way they interpret pain, and one patient's mild pain may be another's moderate pain. The variability within the polychotomous answer key favours an evaluative application, for in order for the questionnaire to pick up a change over time with the dichotomous key, the patient would have to change state from pain to no pain. However, with a polychotomous key, the patient could change from moderate to mild pain. The NHP has a dichotomous item scale and therefore may be particularly relevant for a discriminative application. This was shown by Hilding et al. who used the NHP in Charnley Class A patients and found that the NHP correlated well with RSA results and was, in fact, able to discriminate between patients with continuous migration and stable migration patterns (Hilding et al. 1997). Similar findings using the NHP have been found for hip arthroplasty (Franzen et al. 1997). Admittedly, it has been suggested that other questionnaires would be more appropriate than the NHP (Paper II), but this is when all factors, such as feasibility, patient burden, content validity, and reliability, are considered and equally weighted. The NHP should not be excluded from further possible use, and the findings of Hilding et al. should be investigated further.

In order for a questionnaire to be useful in a discriminative application the total score for a number of items should cover a broad spectrum, for if they all reported the same answers to a questionnaire, discrimination would not be possible. Another way to conceptualize this is that the scores

should follow a normal distribution with little or no ceiling effect. This is why skew and floor and ceiling effect were included in the analyses of Paper II. This rationale may not hold true for an evaluative application. Theoretically, if the results of a questionnaire at time 1 were skewed with a large ceiling effect, meaning that most patients reported the best possible results, then at time 2 it may be easier to evaluate differences if the patients condition worsened. This may be applicable to a post-operative evaluative application of knee arthroplasty, assuming that the natural history is for knee status to deteriorate with time. Although this is theoretically plausible, the variation in evaluative ability based on differences in frequency distribution has not been proven (Liang et al. 1985). Hence, the questionnaires proposed for further use in this thesis in a discriminative fashion may not be the most appropriate for an evaluative application.

Test-retest reliability

As described above, Streiner and Norman suggest that the test-retest reliability of a questionnaire is directly related to the number of items within the questionnaire. Based on this argument, a linear relationship should be evident with respect to number of questionnaire items and ICC value and subsequently, when optimizing for ICC, a researcher would want to choose the questionnaire with the highest number of items. However, the findings in Paper II suggest that this linear model may be an oversimplification. For example, the average ICC value for the 136 item SIP is lower than the ICC value for the 45 item NHP. Also, the 36 item SF-36 has a lower average ICC value than the 12 item SF-12. Finally, the 12 item Oxford-12 has a higher ICC than the 24 item WOMAC. The fact that an inverse relationship to that predicted by Steiner and Norman has been found with these questionnaires may be partly explained by the averaging of ICC values for some questionnaires. Still, the relative number of items per domain is higher in the longer questionnaires; this should not impact significantly on the ICC values. An alternative explanation for this discrepancy may be related more to the variation in item scaling for each questionnaire, such as a simple affirmation for the SIP, versus the dichotomous key for the NHP, versus the polychotomous key for the SF-36.

The problem of noise

All outcome questionnaires tested, ranging from comprehensive general health to a single-item question related directly to the index knee, were influenced by co-morbidity. In some cases this influence was profound with the effect of changing the result of a questionnaire by more than 100%. Other sources of noise that affected some questionnaires, but not all, included gender, age, time since surgery, type of prosthesis (UKA versus TKA) and revision status. Surprisingly, co-morbidity was more significant of a biasing factor than revision status. This has three notable sequelae.

The first sequela relates to the fact that it appears that it is not possible to isolate the knee joint from the person when performing outcomes studies with questionnaires. In this context, person means both the physical and psychosocial self. Failure to isolate the knee joint is consistent, regardless of the complexity or simplicity of the questionnaire tested. Perhaps this evidence refutes the Cartesian Dualism of mind and body proposed by Descartes. Instead, perhaps Pythagoras was correct and man is indeed a measure of all things.

It was contrary to the study hypothesis (Paper IV) to find that the general health questionnaires seemed to be less influenced by co-morbidity, as opposed to even the single-item knee question. Intuitively, a questionnaire that asks about mood, energy level, and body pain with activities of daily living should be more susceptible to the effect of co-morbidity than a single-item questionnaire referring exactly to the index knee. However, it would appear that the general health questionnaires are more stable in this capacity as they are designed specifically for this purpose, that is, to be sensitive to the impact of disease on the physical, mental and social well-being of the person. These findings support the continued use of general health questionnaires in this capacity, at least in association with disease/joint specific questionnaires.

The second sequela is that revision status of the knee may not be an appropriate discriminative index when applying such tests as the ROC Curve. The natural history and the status of revised knees as compared to unrevised knees when measured with outcomes questionnaires simply is not well

enough understood. As the discriminative ability of both a validated and non-validated questionnaire have been compared against revision status using the ROC Curve method more for comparative purposes of the responsiveness of the Oxford-12 against the WOMAC, this should not adversely effect the results and conclusions of this thesis. The natural history of the revised knee is probably not well defined because of the low incidence of knee revision. The material of the SKAR may allow for the natural history to be better delineated, especially with the use of appropriate questionnaires.

The third sequela is the demonstrable requirement for an accounting of co-morbidity when assessing outcomes questionnaires related to knee arthroplasty. However, the stratification of patients by co-morbidity has a deleterious effect on the statistical power of the study. This is compounded when other variables are accounted for, such as age and gender. The large material of the SKAR may allow for adequate power, but further investigation will be necessary. A possible solution to this problem, which may be warranted for small studies, would be to randomize patients entered into the study so that co-morbidity variables would be randomly distributed.

Selecting an outcome questionnaire

The initial thrust of this thesis was to identify questionnaires that would be the most appropriate for use in a wider application to the SKAR. In the process, however, numerous issues have become more apparent with the net effect of clouding the issue of which questionnaires are in fact the best to use. The glib answer to this query can be succinctly stated as "it depends".

Selecting an appropriate questionnaire for a given application is roughly analogous to selecting the appropriate exposure for a camera. In order to select the appropriate exposure, for example, several factors must be accounted for, including the ISO rating of the film, the shutter speed, and the aperture. Essentially, the overall exposure and the final picture result from simultaneously fixing each of these parameters. In doing so, one parameter is optimized in favour of another, with the optimization of said parameter dependent on the desired effect. For example, selecting for a larger aperture allows for a faster shutter speed, but at the expense of depth of field. A lower ISO results in a finer resolution of the picture, but requires a slower shutter speed and or an increased aperture. Conceptually, selecting an appropriate questionnaire is very similar to this process. Just as there is no one exposure setting for a camera, so too is there no one questionnaire suitable for application to all types of health outcomes research.

To select a questionnaire, consideration should first be given to how broadly the desired concepts should be covered, or, more specifically, whether or not a general health, disease specific, site specific, or single-item questionnaire is desired. Choosing a general health questionnaire allows for comparisons to dissimilar groups, which may be of value when assessing the impact on overall health of a given procedure, particularly when compared to another procedure. For example, what is the health value of liver transplantation versus knee arthroplasty?

Once the type of questionnaire is selected, consideration is given to the method in which the questionnaire will be applied—discriminative or evaluative. In a discriminative application, the researcher wants to "freeze the action" and sample the material in an accurate cross-sectional fashion. In an evaluative application, the researchers wishes to observe what happens over time, such as may be seen with a prolonged shutter exposure. In order to optimize for a discriminative application, the questionnaire should have demonstrated good test-retest and internal consistency reliability, and a dichotomous answer scaling is theoretically advantageous. Also, the questionnaire should have a near Normal frequency distribution. For an evaluative application, consideration should be given to demonstrated responsiveness in a setting similar to the proposed study, and polychotomous item scaling may be advantageous. Here, a frequency distribution skewed to the right (preponderance of scores reflecting "good" health status) is favourable, providing that the expected natural history of the measured construct would be for deterioration in health. Previously published responsiveness values should be interpreted with caution because of the profound standardized effect size associated with the surgical intervention for knee arthroplasty patients. Consequently, a published "excellent"

standardized effect size for a questionnaire, when calculated between a pre and post-operative application, may in fact not be responsive enough for an application between 2 post-operative times.

Another consideration when selecting a questionnaire relates to the number of items within the questionnaire. It may be desirable to apply a simple questionnaire to a large number of patients and when response rate is critical, a questionnaire with a low number of items is desirable. The disadvantage of this approach is the loss of detail in the results with a potential for a decrease in the test-retest reliability. In order to increase the amount of sampled detail, more elaborate questionnaires can be selected, but at the expense of response rate.

Finally, the questionnaire to be employed should have been previously validated.

Validation

With the increasing sophistication for health outcomes research in orthopaedics, the use of a "validated questionnaire" has been increasingly called for, as it should, in order to publish results. The validation process is, however, not necessarily rigorous or particularly informative. For example, a rather simple single-item questionnaire has been essentially "validated" in this thesis (Paper III), but this still does not allow the reader to fully comprehend the meaning of the patient's reported satisfaction. Validation is a dynamic process and continued investigation of the performance of a questionnaire across multiple types of application on various cohorts is required. At the expense of the creation of new questionnaires, without compelling reasons, it would be more beneficial to the research community for resources to be directed at continuing this validation process on questionnaires already in use. This would also facilitate future standardization of health outcomes research, at least in orthopaedics.

Future direction

The future direction of this work would involve both a discriminative and evaluative application of questionnaires. Firstly, now that appropriate questionnaires have been identified, a repeat postal survey to all living patients using those questionnaires is feasible. In doing so, subtler variations in outcomes between types of prostheses, resurfaced and non-resurfaced patellae, revised and unrevised knees, etc., may be possible. Also, this repeat application of at least some of the questionnaires would allow for an assessment of the evaluative ability of the questionnaires. The original patient selection could all be resent the original questionnaires that they received, in order to test the evaluative ability of all the questionnaires. As mentioned above, questionnaires other than those chosen as most appropriate may demonstrate favourable status for evaluative use.

It appears that the Modified Charnley Class is a useful and important questionnaire. Future work should concentrate on validating this questionnaire, investigating parameters such as its construct validity, reliability and responsiveness over time.

While the distinct intention of this research was not to develop another outcome questionnaire for knee arthroplasty, it appears that there may be a role to at least reduce the number of items in some questionnaires. Item reduction is plausible, based on the findings in Paper VI and by those of Söderman et al. (2000). Such questionnaire development would require formal psychometric consideration, particularly with respect to reliability. Generally, a reduction in the number of items within a questionnaire adversely affects the reliability (Streiner et al. 1998) but would probably increase the feasibility while reducing patient burden.

The precedence of the discriminative ability of the NHP when correlated to RSA findings is intriguing and worthy of further investigation. It would be worthwhile to test other questionnaires in a similar fashion to see if they maintain the same discriminative ability. Such work is currently ongoing.

National arthroplasty registries are well established in the Nordic countries and are becoming established in other nations, including Canada. It is logically predictable that the current and future proliferation of information technology will motivate and facilitate linking of national registries. Meaningful comparisons between nations using health outcome questionnaires are possible but will be problematic unless several pre-requisites are fulfilled. The first pre-requisite involves standardization of the questionnaires employed. A

consensus between nations is necessary regarding which questionnaires should be used. Obviously, the agreed upon questionnaires should be available in a translated and subsequently validated version for the respective nations. Several types of questionnaires should be agreed upon in order to optimize for the specific applications, as outlined above. The second pre-requisite involves the establishment of demographic norms for each nation. Such norms would provide the required "denominator" in order to compare outcomes results. The final pre-requisite involves more detailed subjective descriptions of the natural history of knee arthroplasty, including the effect of complications. Currently, the natural history is not well described, hence making the creation of hypothetical constructs difficult. For example, this thesis used revised versus unrevised patients as the construct for comparing the discriminative ability of the WOMAC and Oxford-12 using the ROC Curve method. As the natural subjective history for both revised and unrevised knee arthroplasty patients is not well defined, so too is the construct. This weakens the test results. The work of the SKAR and other national registries will be instrumental in this capacity.

Conclusions

1. A large-scale postal survey is feasible to knee arthroplasty patients in Sweden. High usable response rates and low patient burden can be expected with most relevant questionnaires.

2. The SF-12 and the Oxford-12 Item Knee Score appeared to be the most appropriate general health and disease/site specific questionnaires, respectively, for use in a large-scale postal survey in a cross-sectional fashion when considering feasibility, patient burden, content validity and reliability.

3. Global single-item questionnaires can yield discriminative data when applied to the SKAR, such as variations in patient satisfaction between revised and unrevised knees. Usable response rates are higher for the single-item questionnaires but reliability is lower.

4. Generally, when patients state that they are satisfied after knee arthroplasty, they are referring to relief of pain primarily and improved function secondarily.

5. All questionnaires tested in this thesis were strongly biased by patient co-morbidity, as measured by a modified Charley Class for knee arthroplasty. Co-morbidity should be accounted for when evaluating the results of arthroplasty. It appears that it is not possible to isolate the knee from the body and the mind with health outcomes questionnaires.

6. Patients who score one standard deviation worse than the mean pre-operative WOMAC Pain and Physical Function domains scores do not reach the same 1 year post-operative status as patients scoring one standard deviation better than the mean score.

7. The Swedish translated version of the Oxford-12 Item Knee Score is linguistically and culturally equivalent to the English version and has acceptable psychometric characteristics in keeping with the original questionnaire. The validation process should continue.

Acknowledgements

I have been extremely fortunate to have the academic freedom to complete this thesis. I would like to express my gratitude to the following people for their various roles in seeing this thesis to fruition.

- **Leif Ryd**, my supervisor, mentor, and friend, for teaching me how to be a thoughtful researcher, for stimulating philosophical conversation (especially over sushi), and for making all this possible.

- **Lars Lidgren**, my co-supervisor and Professor, for the privilege of studying in Lund and for strong leadership. But especially for the foresight of arranging for me to work with Otto!

- **Otto Robertsson**, my co-author and friend, for grounding my flighty ideas with healthy doses of reality, for expert instruction on databases, for brilliant insights, and for ex-patriot comradery.

- **Thorbjörn Pehrsson**, for easing me into Swedish culture and for being a true friend.

- **Jonas Ranstam**, for expert statistical advice.

- **Michael Gross**, my colleague and friend, for starting me off on this pathway of academic pursuits (without my implicit knowledge, I might add).

- **Robert Bourne and Cecil Rorabeck**, for the privilege of working as their fellow, for academic insight, and for showing me by example that academic and clinical excellence is possible in the Canadian health care environment.

- **Gun-Britt Nyberg**, for kindly assisting me with all my administrative type problems which made my stay in Sweden possible and far less stressful.

- **Kaj Knutson**, for typesetting and editing this work.

- The knee arthroplasty patients of Sweden, for tireless efforts in filling out questionnaires without complaint.

- The participating clinics in Sweden without whom this thesis, and the SKAR, would not be possible.

- My colleagues in the Biomek Lab, for support, friendship, making me feel welcome and especially for hackey sack.

- **Alexander Lidgren**, for friendship.

- **My parents**, for making everything possible with unconditional support and encouragement.

- **Monica**, my wife, for support, encouragement, and allowing me the freedom to chase after my dreams.

- **Jon**, my son, for making it all worthwhile.

The studies within this thesis were supported by grants from Stiftelsen för bistånd åt rörelsehindrade i Skåne, Greta och Johan Kocks Stiftelse, Österlunds Stiftelse, Konung Gustaf V:s 80-årsfond, Socialstyrelsen/Landstingsförbundet, the Swedish Medical Research Council (Project 9509), the Arthritis Society of Canada and the Medical Faculty of the University of Lund.

References

Aichroth P, Freeman M A R, Smillie I S, Souter W A. A knee function assessment chart. From the British Orthopaedic Association Research Sub-Committee. J Bone Joint Surg [Br] 1978; 60-B (3): 308-9.

Amendola A, Rorabeck C H, Bourne R B, Apyan P M. Total knee arthroplasty following high tibial osteotomy for osteoarthritis. J Arthroplasty 1989; 4 (Suppl): S11-7.

Anderson J G, Wixson R L, Tsai D, Stulberg S D, Chang R W. Functional outcome and patient satisfaction in total knee patients over the age of 75. J Arthroplasty 1996; 11 (7): 831-40.

Apley A G. An assessment of assessment [editorial]. J Bone Joint Surg [Br] 1990; 72 (6): 957-8.

Armstrong R A, Whiteside L A. Results of cementless total knee arthroplasty in an older rheumatoid arthritis population. J Arthroplasty 1991; 6 (4): 357-62.

Asch D A, Christakis N A. Different response rates in a trial of two envelop styles in mail survey research. Epidemiology 1994; 5 (3): 364-5.

Barrack R L, Smith P, Munn B, Engh G, Rorabeck C. The Ranawat Award. Comparison of surgical approaches in total knee arthroplasty. Clin Orthop 1998; (356): 16-21.

Barrack R L, Wolfe M W, Waldman D A, Milicic M, Bertot A J, Myers L. Resurfacing of the patella in total knee arthroplasty. A prospective, randomized, double-blind study . J Bone Joint Surg [Am] 1997; 79 (8): 1121-31.

Beaton D E, Hogg-Johnson S, Bombardier C. Evaluating changes in health status: reliability and responsiveness of five generic health status measures in workers with musculoskeletal disorders. J Clin Epidemiol 1997; 50 (1): 79-93.

Bellamy N. Pain assessment in osteoarthritis: experience with the WOMAC osteoarthritis index. Semin Arthritis Rheum 1989; 18 (4 Suppl 2): 14-7.

Bellamy N. Outcome measurement in osteoarthritis clinical trials. J Rheumatol Suppl 1995; 43 : 49-51.

Bellamy N, Buchanan W W. Outcome measurement in osteoarthritis clinical trials: the case for standardisation. Clin Rheumatol 1984; 3 (3): 293-303.

Bellamy N, Buchanan W W, Goldsmith C H, Campbell J, Stitt L W. Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. J Rheumatol 1988; 15 (12): 1833-40.

Bellamy N, Goldsmith C H, Buchanan W W, Campbell J, Duku E. Prior score availability: observations using the WOMAC osteoarthritis index [letter]. Br J Rheumatol 1991; 30 (2): 150-1.

Bellamy N, Kean W F, Buchanan W W, Gerecz-Simon E, Campbell J. Double blind randomized controlled trial of sodium meclofenamate (Meclomen) and diclofenac sodium (Voltaren): post validation reapplication of the WOMAC Osteoarthritis Index. J Rheumatol 1992; 19 (1): 153-9.

Bengtson S, Knutson K. The infected knee arthroplasty. A 6-year follow-up of 357 cases. Acta Orthop Scand 1991; 62 (4): 301-11.

Bengtson S, Knutson K, Lidgren L. Revision of infected knee arthroplasty. Acta Orthop Scand 1986; 57 (6): 489-94.

Bengtson S, Knutson K, Lidgren L. Treatment of infected knee arthroplasty. Clin Orthop 1989; (245): 173-8.

Bergner M, Bobbitt R A, Carter W B, Gilson B S. The Sickness Impact Profile: development and final revision of a health status measure. Med Care 1981; 19 (8): 787-805.

Bland J M, Altman D G. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1 (8476): 307-10.

Bland J M, Altman D G. Measurement error and correlation coefficients. BMJ 1996; 313 (7048): 41-2.

Bland J M, Altman D G. Cronbach's alpha. BMJ 1997; 314 (7080): 572.

Bland M. An Introduction to Medical Statistics. Oxford University Press, New York, 1995.

Bombardier C, Melfi C A, Paul J, Green R, Hawker G, Wright J, Coyte P. Comparison of a generic and a disease-specific measure of pain and physical function after knee replacement surgery. Med Care 1995; 33 (4 Suppl): AS131-44.

Braeken A M, Lochhaas-Gerlach J A, Gollish J D, Myles J D, Mackenzie T A. Determinants of 6-12 month postoperative functional status and pain after elective total hip replacement . Int J Qual Health Care 1997; 9 (6): 413-8.

Brazier J, Jones N, Kind P. Testing the validity of the Euroqol and comparing it with the SF-36 health survey questionnaire . Qual Life Res 1993; 2 (3): 169-80.

Brazier J E, Harper R, Jones N M, O'Cathain A, Thomas K J, Usherwood T, Westlake L. Validating the SF-36 health survey questionnaire: new outcome measure for primary care . BMJ 1992; 305 (6846): 160-4.

Brinker M R, Lund P J, Barrack R L. Demographic biases of scoring instruments for the results of total knee arthroplasty. J Bone Joint Surg [Am] 1997; 79 (6): 858-65.

Brown L. The New Shorter Oxford English Dictionary. Oxford University Press, New York, 1993.

Callahan C M, Drake B G, Heck D A, Dittus R S. Patient outcomes following tricompartmental total knee replacement. A meta-analysis. JAMA 1994; 271 (17): 1349-57.

Centor R M. Signal detectability: the use of ROC curves and their analyses. Med Decis Making 1991; 11 (2): 102-6.

Charnley J. Low Friction Arthroplasty of the Hip. Springer-Verlag, Berlin, 1979.

Cronbach L J, Meehl, P.E. Construct validity in psychological tests. Psych Bull 1955; 52 : 281-302.

Damiano A M. Sickness Impact Profile user's manual and interpretation guide. John Hopkins University Press, Baltimore, 1996.

Dawson J, Fitzpatrick R, Carr A. Questionnaire on the perceptions of patients about shoulder surgery. J Bone Joint Surg [Br] 1996a; 78 (4): 593-600.

Dawson J, Fitzpatrick R, Carr A, Murray D. Questionnaire on the perceptions of patients about total hip replacement . J Bone Joint Surg [Br] 1996b; 78 (2): 185-90.

Dawson J, Fitzpatrick R, Murray D, Carr A. The problem of "noise" in monitoring patient-based outcomes: generic, disease-specific and site-specific instruments for total hip replacement. J Health Serv Res Policy 1996c; 1 (4): 224-231.

Dawson J, Fitzpatrick R, Murray D, Carr A. Questionnaire on the perceptions of patients about total knee replacement. J Bone Joint Surg [Br] 1998; 80 (1): 63-9.

Descartes R. A Discourse on Method: Meditations and Principles. Guernsey Press Company Limited, Guernsey, 1986.

Deyo R A, Centor R M. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. J Chronic Dis 1986; 39 (11): 897-906.

Deyo R A, Inui T S, Leininger J D, Overman S S. Measuring functional outcomes in chronic disease: a comparison of traditional scales and a self-administered health status questionnaire in patients with rheumatoid arthritis. Med Care 1983; 21 (2): 180-92.

Di Fabio R P, Boissonnault W. Physical therapy and health-related outcomes for patients with common orthopaedic diagnoses. J Orthop Sports Phys Ther 1998; 27 (3): 219-30.

Dolan P, Kind P. Inconsistency and health state valuations. Soc Sci Med 1996; 42 (4): 609-15.

Drake B G, Callahan C M, Dittus R S, Wright J G. Global rating systems used in assessing knee arthroplasty outcomes. J Arthroplasty 1994; 9 (4): 409-17.

Engelberg R, Martin D P, Agel J, Obremsky W, Coronado G, Swiontkowski M F. Musculoskeletal Function Assessment instrument: criterion and construct validity. J Orthop Res 1996; 14 (2): 182-92.

Engelbrecht E. Sliding prosthesis, a partial prosthesis in destructive processes of the knee joint. Chirurg 1971; 42 (11): 510-4.

Essink-Bot M L, Krabbe P F, Bonsel G J, Aaronson N K. An empirical comparison of four generic health status measures. The Nottingham Health Profile, the Medical Outcomes Study 36-item Short- Form Health Survey, the COOP/WONCA charts, and the EuroQol instrument. Med Care 1997; 35 (5): 522-37.

Fehring T K, Valadie A L. Knee instability after total knee arthroplasty. Clin Orthop 1994; (299): 157-62.

Feinstein A R. Clinimetrics. Yale University Press, New Haven, 1987.

Franzen H, Johnsson R, Nilsson L T. Impaired quality of life 10 to 20 years after primary hip arthroplasty. J Arthroplasty 1997; 12 (1): 21-4.

Freeman M A, Levack B. British contribution to knee arthroplasty. Clin Orthop 1986; (210): 69-79.

Gandek B, Ware J E, Aaronson N K, Apolone G, Bjorner J B, Brazier J E, Bullinger M, Kaasa S, Leplege A, Prieto L, Sullivan M. Cross-validation of item selection and scoring for the SF-12 Health Survey in nine countries: results from the IQOLA Project. International Quality of Life Assessment. J Clin Epidemiol 1998; 51 (11): 1171-8.

Garellick G, Malchau H, Herberts P. Specific or general health outcome measures in the evaluation of total hip replacement. A comparison between the Harris hip score and the Nottingham Health Profile. J Bone Joint Surg [Br] 1998; 80 (4): 600-6.

Gluck T. Die Invaginationsmethode der Osteo- und Arthroplastick. Berl Klin Wschr 1890; 19: 732.

Gross M. A critique of the methodologies used in clinical studies of hip-joint arthroplasty published in the English-language orthopaedic literature. J Bone Joint Surg [Am] 1988; 70 (9): 1364-71.

Guillemin F, Bombardier C, Beaton D. Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. J Clin Epidemiol 1993; 46 (12): 1417-32.

Gunston F H. Polycentric knee arthroplasty. Prosthetic simulation of normal knee movement. J Bone Joint Surg [Br] 1971; 53 (2): 272-7.

Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chronic Dis 1987; 40 (2): 171-8.

Guyatt G H. The philosophy of health-related quality of life translation. Qual Life Res 1993; 2 (6): 461-5.

Hanley J A, McNeil B J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143 (1): 29-36.

Hawker G, Melfi C, Paul J, Green R, Bombardier C. Comparison of a generic (SF-36) and a disease specific (WOMAC) (Western Ontario and McMaster Universities Osteoarthritis Index) instrument in the measurement of outcomes after knee replacement surgery. J Rheumatol 1995; 22 (6): 1193-6.

Hawker G, Wright J, Coyte P, Paul J, Dittus R, Croxford R, Katz B, Bombardier C, Heck D, Freund D. Health-related quality of life after knee replacement. J Bone Joint Surg [Am] 1998; 80 (2): 163-73.

Hays R D, Anderson R, Revicki D. Psychometric considerations in evaluating health-related quality of life measures. Qual Life Res 1993; 2 (6): 441-9.

Heck D A, Robinson R L, Partridge C M, Lubitz R M, Freund D A. Patient outcomes after knee replacement. Clin Orthop 1998; (356): 93-110.

Hilding M B, Bäckbro B, Ryd L. Quality of life after knee arthroplasty. A randomized study of 3 designs in 42 patients, compared after 4 years. Acta Orthop Scand 1997; 68 (2): 156-60.

Hill A, Roberts J, Ewings P, Gunnell D. Non-response bias in a lifestyle survey. J Public Health Med 1997; 19 (2): 203-7.

Hirsch H S, Lotke P A, Morrison L D. The posterior cruciate ligament in total knee surgery. Save, sacrifice, or substitute? Clin Orthop 1994; (309): 64-8.

Hoher J, Bach T, Munster A, Bouillon B, Tiling T. Does the mode of data collection change results in a subjective knee score? Self-administration versus interview. Am J Sports Med 1997; 25 (5): 642-7.

Hunt S M, McKenna S P, McEwen J, Backett E M, Williams J, Papp E. A quantitative approach to perceived health status: a validation study. J Epidemiol Community Health 1980; 34 (4): 281-6.

Hunt S M, McKenna S P, McEwen J, Williams J, Papp E. The Nottingham Health Profile: subjective health status and medical consultations. Soc Sci Med [A] 1981a; 15: 221-9.

Hunt S M, McKenna S P, Williams J. Reliability of a population survey tool for measuring perceived health problems: a study of patients with osteoarthrosis. J Epidemiol Community Health 1981b; 35 (4): 297-300.

Hurst N P, Kind P, Ruta D, Hunter M, Stubbings A. Measuring health-related quality of life in rheumatoid arthritis: validity, responsiveness and reliability of EuroQol (EQ-5D). Br J Rheumatol 1997; 36 (5): 551-9.

Insall J N, Ranawat C S, Aglietti P, Shine J. A comparison of four models of total knee-replacement prostheses. J Bone Joint Surg [Am] 1976; 58 (6): 754-65.

Insall J, Scott W N, Ranawat C S. The total condylar knee prosthesis. A report of two hundred and twenty cases. J Bone Joint Surg [Am] 1979; 61 (2): 173-80.

Insall J N, Dorr L D, Scott R D, Scott W N. Rationale of the Knee Society clinical rating system. Clin Orthop 1989; (248): 13-4.

Jenkinson C, Layte R, Jenkinson D, Lawrence K, Petersen S, Paice C, Stradling J. A shorter form health survey: can the SF-12 replicate results from the SF-36 in longitudinal studies? J Public Health Med 1997; 19 (2): 179-86.

Jenkinson C, Wright L, Coulter A. Criterion validity and reliability of the SF-36 in a population sample. Qual Life Res 1994; 3 (1): 7-12.

Jette A M. The Functional Status Index: reliability and validity of a self-report functional disability measure. J Rheumatol 1987; 14 Suppl 15 : 15-21.

Jette A M, Davies A R, Cleary P D, Calkins D R, Rubenstein L V, Fink A, Kosecoff J, Young R T, Brook R H, Delbanco T L. The Functional Status Questionnaire: reliability and validity when used in primary care [published erratum appears in J Gen Intern Med 1986 Nov-Dec;1(6):427]. J Gen Intern Med 1986; 1 (3): 143-9.

Joseph J, Kaufman E E. Preliminary results of Miller-Galante uncemented total knee arthroplasty. Orthopedics 1990; 13 (5): 511-6.

Kaplan R M, Bush J W, Berry C C. Health status: types of validity and the index of well-being. Health Serv Res 1976; 11 (4): 478-507.

Katz J N, Wright E A, Guadagnoli E, Liang M H, Karlson E W, Cleary P D. Differences between men and women undergoing major orthopedic surgery for degenerative arthritis. Arthritis Rheum 1994; 37 (5): 687-94.

Kazis L E, Anderson J J, Meenan R F. Effect sizes for interpreting changes in health status. Med Care 1989; 27 (3 Suppl): S178-89.

Kinnersley P, Peters T, Stott N. Measuring functional health status in primary care using the COOP-WONCA charts: acceptability, range of scores, construct validity, reliability and sensitivity to change. Br J Gen Pract 1994; 44 (389): 545-9.

Kirshner B, Guyatt G. A methodological framework for assessing health indices. J Chronic Dis 1985; 38 (1): 27-36.

Knight J L, Atwater R D, Grothaus L. Clinical results of the modular porous-coated anatomic (PCA) total knee arthroplasty with cement: a 5-year prospective study. Orthopedics 1997; 20 (11): 1025-33.

Knutson K, Hovelius L, Lindstrand A, Lidgren L. Arthrodesis after failed knee arthroplasty. A nationwide multicenter investigation of 91 cases. Clin Orthop 1984; (191): 202-11.

Knutson K, Lewold S, Robertsson O, Lidgren L. The Swedish knee arthroplasty register. A nation-wide study of 30,003 knees 1976-1992. Acta Orthop Scand 1994; 65 (4): 375-86.

Knutson K, Lindstrand A, Lidgren L. Survival of knee arthroplasties. A nation-wide multicentre investigation of 8000 cases. J Bone Joint Surg [Br] 1986; 68 (5): 795-803.

Knutson K, Tjörnstrand B, Lidgren L. Survival of knee arthroplasties for rheumatoid arthritis. Acta Orthop Scand 1985; 56 (5): 422-5.

Konig A, Scheidler M, Rader C, Eulert J. The need for a dual rating system total knee arthroplasty. Clin Orthop 1997; (345): 161-7.

Kreibich D N, Vaz M, Bourne R B, Rorabeck C H, Kim P, Hardie R, Kramer J, Kirkley A. What is the best way of assessing outcome after total knee replacement? Clin Orthop 1996; (331): 221-5.

Laupacis A, Bourne R, Rorabeck C, Feeny D, Wong C, Tugwell P, Leslie K, Bullas R. The effect of elective total hip replacement on health-related quality of life. J Bone Joint Surg [Am] 1993; 75 (11): 1619-26.

Lequesne M G, Mery C, Samson M, Gerard P. Indexes of severity for osteoarthritis of the hip and knee. Validation— value in comparison with other assessment tests [published errata appear in Scand J Rheumatol Suppl 1988;73:1 and Scand J Rheumatol 1988;17(3):following 241]. Scand J Rheumatol Suppl 1987; 65 : 85-9.

Lequesne M. Informational indices. Validation of criteria and tests. Scand J Rheumatol Suppl 1989; 80 : 17-28.

Lequesne M G, Samson M. Indices of severity in osteoarthritis for weight bearing joints. J Rheumatol Suppl 1991; 27 : 16-8.

Lequesne M, Fannius J, Reginster J Y, Verdickt W, du Laurier M V. Floctafenin versus acetaminophen for pain control in patients with osteoarthritis in the lower limbs. Franco-Belgian Task Force. Rev Rhum Engl Ed 1997a; 64 (5): 327-33.

Lequesne M G. The algofunctional indices for hip and knee osteoarthritis. J Rheumatol 1997b; 24 (4): 779-81.

Lescoe-Long M, Long M J, Johnston D W. Quality of life improvement in patients with osteoarthritis: the potential for office-based assessment. Arthritis Care Res 1996; 9 (3): 177-81.

Lewold S, Knutson K, Lidgren L. Reduced failure rate in knee prosthetic surgery with improved implantation technique. Clin Orthop 1993; (287): 94-7.

Lewold S, Olsson H, Gustafson P, Rydholm A, Lidgren L. Overall cancer incidence not increased after prosthetic knee replacement: 14,551 patients followed for 66,622 person-years. Int J Cancer 1996; 68 (1): 30-3.

Lewold S, Robertsson O, Knutson K, Lidgren L. Revision of unicompartmental knee arthroplasty: outcome in 1,135 cases from the Swedish Knee Arthroplasty study. Acta Orthop Scand 1998; 69 (5): 469-74.

Liang M H, Fossel A H, Larson M G. Comparisons of five health status instruments for orthopedic evaluation. Med Care 1990; 28 (7): 632-42.

Liang M H, Larson M G, Cullen K E, Schwartz J A. Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. Arthritis Rheum 1985; 28 (5): 542-7.

Lieberman J R, Dorey F, Shekelle P, Schumacher L, Kilgus D J, Thomas B J, Finerman G A. Outcome after total hip arthroplasty. Comparison of a traditional disease-specific and a quality-of-life measurement of outcome. J Arthroplasty 1997; 12 (6): 639-45.

Lohmander L S, Dalen N, Englund G, Hämälainen M, Jensen E M, Karlsson K, Odensten M, Ryd L, Sernbo I, Suomalainen O, Tegnander A. Intra-articular hyaluronan injections in the treatment of osteoarthritis of the knee: a randomised, double blind, placebo controlled multicentre trial. Hyaluronan Multicentre Trial Group. Ann Rheum Dis 1996; 55 (7): 424-31.

MacDonagh R P, Cliff A M, Speakman M J, O'Boyle P J, Ewings P, Gudex C. The use of generic measures of health-related quality of life in the assessment of outcome from transurethral resection of the prostate. Br J Urol 1997; 79 (3): 401-8.

Marmor L. The modular knee. Clin Orthop 1973; 94: 242-8.

Martin D P, Engelberg R, Agel J, Snapp D, Swiontkowski M F. Development of a musculoskeletal extremity health status instrument: the Musculoskeletal Function Assessment instrument. J Orthop Res 1996; 14 (2): 173-81.

Martin D P, Engelberg R, Agel J, Swiontkowski M F. Comparison of the Musculoskeletal Function Assessment questionnaire with the Short Form-36, the Western Ontario and McMaster Universities Osteoarthritis Index, and the Sickness Impact Profile health-status measures. J Bone Joint Surg [Am] 1997; 79 (9): 1323-35.

Mathias S D, Fifer S K, Patrick D L. Rapid translation of quality of life measures for international clinical trials: avoiding errors in the minimalist approach. Qual Life Res 1994; 3 (6): 403-12.

McHorney C A, Ware J E, Jr., Rogers W, Raczek A E, Lu J F. The validity and relative precision of MOS short- and long-form health status scales and Dartmouth COOP charts. Results from the Medical Outcomes Study. Med Care 1992; 30 (5 Suppl): MS253-65.

McHorney C A, Kosinski M, Ware J E, Jr. Comparisons of the costs and quality of norms for the SF-36 health survey collected by mail versus telephone interview: results from a national survey. Med Care 1994a; 32 (6): 551-67.

McHorney C A, Ware J E, Jr., Lu J F, Sherbourne C D. The MOS 36-item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. Med Care 1994b; 32 (1): 40-66.

McHorney C A. Generic health measurement: past accomplishments and a measurement paradigm for the 21st century. Ann Intern Med 1997; 127 (8 Pt 2): 743-50.

Meenan R F, Kazis L E, Anthony J M, Wallin B A. The clinical and health status of patients with recent-onset rheumatoid arthritis. Arthritis Rheum 1991; 34 (6): 761-5.

Miller A, Friedman B. Fascial Arthroplasty of the knee. J Bone Joint Surg [Am] 1952; 34-A: 55.

Nafei A, Kristensen O, Kjaersgaard-Andersen P, Hvid I, Jensen J. Total condylar arthroplasty for gonarthrosis. A prospective 10-year study of 138 primary cases. Acta Orthop Scand 1993; 64 (4): 421-7.

Nilsson L T, Franzen H, Carlsson A S, Önnerfält R. Early radiographic loosening impairs the function of a total hip replacement. The Nottingham Health Profile of 49 patients at five years. J Bone Joint Surg [Br] 1994; 76 (2): 235-9.

Nunnally J C, Bernstein I H. Psychometric Theory. McGraw-Hill, New York, 1994.

Palmer R H. Process-based measures of quality: the need for detailed clinical data in large health care databases. Ann Intern Med 1997; 127 (8 Pt 2): 733-8.

Patrick D L, Deyo R A. Generic and disease-specific measures in assessing health status and quality of life. Med Care 1989; 27 (3 Suppl): S217-32.

Plant P, McEwen J, Prescott K. Use of the Nottingham Health Profile to test the validity of census variables to proxy the need for health care. J Public Health Med 1996; 18 (3): 313-20.

Pollard W E, Bobbitt R A, Bergner M, Martin D P, Gilson B S. The Sickness Impact Profile: reliability of a health status measure. Med Care 1976; 14 (2): 146-55.

Ries M D, Philbin E F, Groff G D, Sheesley K A, Richman J A, Lynch F, Jr. Improvement in cardiovascular fitness after total knee arthroplasty. J Bone Joint Surg [Am] 1996; 78 (11): 1696-701.

Rissanen P, Aro S, Slätis P, Sintonen H, Paavolainen P. Health and quality of life before and after hip or knee arthroplasty. J Arthroplasty 1995; 10 (2): 169-75.

Ritter M A, Albohm M J, Keating E M, Faris P M, Meding J B. Comparative outcomes of total joint arthroplasty. J Arthroplasty 1995; 10 (6): 737-41.

Robertsson O, Knutson K, Lewold S, Goodman S, Lidgren L. Knee arthroplasty in rheumatoid arthritis. A report from the Swedish Knee Arthroplasty Register on 4,381 primary operations 1985-1995. Acta Orthop Scand 1997; 68 (6): 545-53.

Robertsson O, Borgquist L, Knutson K, Lewold S, Lidgren L. Use of unicompartmental instead of tricompartmental prostheses for unicompartmental arthrosis in the knee is a cost-effective alternative. 15,437 primary tricompartmental prostheses were compared with 10,624 primary medial or lateral unicompartmental prostheses. Acta Orthop Scand 1999a; 70 (2): 170-5.

Robertsson O, Dunbar M, Knutson K, Lewold S, Lidgren L. Validation of the Swedish Knee Arthroplasty Register: a postal survey regarding 30,376 knees operated on between 1975 and 1995. Acta Orthop Scand 1999b; 70 (5): 467-72.

Robertsson O, Dunbar M J, Knutson K, Lewold S, Lidgren L. The Swedish Knee Arthroplasty Register. 25 years experience. Bull Hosp Jt Dis 1999c; 58 (3): 133-8.

Robertsson O, Knutson K, Lewold S, Lidgren L. Knee Arthroplasty for Osteoarthrosis and Rheumatoid Arthritis 1986-1996. Scientific Exhibit SE028, Annual Meeting of the American Academy of Orthopaedic Surgeons, Anaheim 1999d.

Robertsson O, The Swedish Knee Arthroplasty Register: Validity and Outcome. Thesis, Lund University, Lund, Sweden 2000.

Robertsson O, Dunbar M J, Knutson K, Lidgren L. Past incidence and future demand for knee arthroplasty in Sweden: a report from the Swedish Knee Arthroplasty Register regarding the effect of past and future population changes on the number of arthroplasties performed. Acta Orthop Scand 2000; 71 (4): 376-80.

Roos E M, Klässbo M, Lohmander L S. WOMAC osteoarthritis index: reliability, validity, and responsiveness in patients with arthroscopically assessed osteoarthritis. Scand J Rheumatol 1999; 53(9): 716-21

Ryd L, Kärrholm J, Ahlvin P. Knee scoring systems in gonarthrosis. Evaluation of interobserver variability and the envelope of bias. Score Assessment Group . Acta Orthop Scand 1997; 68 (1): 41-5.

Schrøder H M, Kristensen P W, Petersen M B, Nielsen P T. Patient survival after total knee arthroplasty. 5-year data in 926 patients. Acta Orthop Scand 1998; 69 (1): 35-8.

Schroeder-Boersch H, Scheller G, Fischer J, Jani L. Advantages of patellar resurfacing in total knee arthroplasty. Two-year results of a prospective randomized study. Arch Orthop Trauma Surg 1998; 117 (1-2): 73-8.

Shiers L G. Arthroplasty of the Knee. Preliminary report of new method. J Bone Joint Surg [Br] 1954; 36: 553-60.

Söderman P, Malchau H, Herberts P. Outcome after total hip arthroplasty: Part I. General health evaluation in relation to definition of failure in the Swedish National Total Hip Arthoplasty register. Acta Orthop Scand 2000; 71 (4): 354-9.

Speed J S, Trout P C. Arthroplasty of the Knee. A follow-up study. J Bone Joint Surg [Br] 1949; 31-B: 53.

Streiner D L, Norman G R. Health Measurement Scales: A Practical Guide to their Development and Use, Oxford University Press Inc., New York, 1998.

Strohmeier J, Westbrook P. Divine Harmony: The Life and Teachings of Pythagoras, Berkely Hill Books, Berkeley, 1999.

Stucki G, Liang M H, Phillips C, Katz J N. The Short Form-36 is preferable to the SIP as a generic health status measure in patients undergoing elective total hip arthroplasty. Arthritis Care Res 1995; 8 (3): 174-81.

Sullivan M. Sickness impact profile: introduction of a Swedish version of health status indicators. Läkartidningen 1985; 82 (20): 1861-2.

Sullivan M. Measuring quality of life. A new general and a new tumor specific formulary for evaluation and planning. Läkartidningen 1994; 91 (13): 1340-1.

Sullivan M, Ahlmen M, Archenholtz B, Svensson G. Measuring health in rheumatic disorders by means of a Swedish version of the sickness impact profile. Results from a population study. Scand J Rheumatol 1986; 15 (2): 193-200.

Sullivan M, Karlsson J, Ware J E, Jr. The Swedish SF-36 Health Survey—I. Evaluation of data quality, scaling assumptions, reliability and construct validity across general populations in Sweden. Soc Sci Med 1995; 41 (10): 1349-58.

Sun Y, Sturmer T, Gunther K P, Brenner H. Reliability and validity of clinical outcome measurements of osteoarthritis of the hip and knee—a review of the literature. Clin Rheumatol 1997; 16 (2): 185-98.

Tew M, Waugh W. Estimating the survival time of knee replacement. J Bone Joint Surg [Br] 1982; 64 (5): 579-82.

Thompson S P. Life of Lord Kelvin. Chelsea Publishing, Chelsea, 1976.

Walldius B. Arthroplasty of the knee using an endoprosthesis [classical article]. Clin Orthop 1996; (331): 4-10.

Ware J, Jr., Kosinski M, Keller S D. A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. Med Care 1996; 34 (3): 220-33.

Ware J E, Jr., Sherbourne C D. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992; 30 (6): 473-83.

Wiklund I, Dimenas E. The Swedish version of the Nottingham Health Profile. A questionnaire for the measurement of health-related quality of life. Läkartidningen 1990; 87 (18): 1575-6.

Wiklund I, Romanus B. A comparison of quality of life before and after arthroplasty in patients who had arthrosis of the hip joint. J Bone Joint Surg [Am] 1991; 73 (5): 765-9.

Wiklund I, Romanus B, Hunt S M. Self-assessed disability in patients with arthrosis of the hip joint. Reliability of the Swedish version of the Nottingham Health Profile. Int Disabil Stud 1988; 10 (4): 159-63.

Williams J I, Llewellyn Thomas H, Arshinoff R, Young N, Naylor C D. The burden of waiting for hip and knee replacements in Ontario. Ontario Hip and Knee Replacement Project Team. J Eval Clin Pract 1997; 3 (1): 59-68.

Wolfe F, Hawley D J. Measurement of the quality of life in rheumatic disorders using the EuroQol. Br J Rheumatol 1997; 36 (7): 786-93.

Wright J G, Young N L. A comparison of different indices of responsiveness. J Clin Epidemiol 1997; 50 (3): 239-46.

Appendix (Oxford-12)

PROBLEMS WITH YOUR KNEE

During the past 4 weeks.. ✓tick one box for every question

1 *During the past 4 weeks.....*
How would you describe the pain you usually have from your knee?

None	Very mild	Mild	Moderate	Severe
❑	❑	❑	❑	❑

2 *During the past 4 weeks.....*
Have you had any trouble with washing and drying yourself (all over) because of your knee?

No trouble at all	Very little trouble	Moderate trouble	Extreme difficulty	Impossible to do
❑	❑	❑	❑	❑

3 *During the past 4 weeks.....*
Have you had any trouble getting in and out of a car or using public transport because of your knee? (whichever you would tend to use)

No trouble at all	Very little trouble	Moderate trouble	Extreme difficulty	Impossible to do
❑	❑	❑	❑	❑

4 *During the past 4 weeks.....*
For how long have you been able to walk before pain from your knee becomes **severe**? (*with or without a stick*)

No pain/ More than 30 minutes	16 to 30 minutes	5 to 15 minutes	Around the house only	Not at all - pain severe when walking
❑	❑	❑	❑	❑

5 *During the past 4 weeks.....*
After a meal (sat at a table), how painful has it been for you to stand up from a chair because of your knee?

Not at all painful	Slightly painful	Moderately painful	Very painful	Unbearable
❑	❑	❑	❑	❑

6 *During the past 4 weeks.....*
Have you been limping when walking, because of your knee?

Rarely/ never	Sometimes, or just at first	Often, not just at first	Most of the time	All of the time
❑	❑	❑	❑	❑

Oxford Knee Score Department of Public Health, Institute of Health Sciences, Old Road, Oxford OX3 7LF

/P.T.O

During the past 4 weeks...

✓tick one box for every question

7 *During the past 4 weeks.....*

Could you kneel down and get up again afterwards?

Yes, Easily	With little difficulty	With moderate difficulty	With extreme difficulty	No, Impossible
❑	❑	❑	❑	❑

8 *During the past 4 weeks.....*

Have you been troubled by pain from your knee in bed at night?

No nights	Only 1 or 2 nights	Some nights	Most nights	Every night
❑	❑	❑	❑	❑

9 *During the past 4 weeks.....*

How much has pain from your knee interfered with your usual work *(including housework)*?

Not at all	A little bit	Moderately	Greatly	Totally
❑	❑	❑	❑	❑

10 *During the past 4 weeks.....*

Have you felt that your knee might suddenly 'give way' or let you down?

Rarely/ never	Sometimes, or just at first	Often, not just at first	Most of the time	All of the time
❑	❑	❑	❑	❑

11 *During the past 4 weeks.....*

Could you do the household shopping on your own?

Yes, Easily	With little difficulty	With moderate difficulty	With extreme difficulty	No, Impossible
❑	❑	❑	❑	❑

12 *During the past 4 weeks.....*

Could you walk down one flight of stairs?

Yes, Easily	With little difficulty	With moderate difficulty	With extreme difficulty	No, Impossible
❑	❑	❑	❑	❑

Problem med ditt knä

Under de senaste fyra veckorna...

Markera en ruta för varje fråga

Under de senaste fyra veckorna...

1 Hur skulle Du beskriva den smärta Du vanligtvis har i Ditt knä?

Ingen	Mycket lindrig	Lindrig	Måttlig	Svår
☐	☐	☐	☐	☐

Under de senaste fyra veckorna...

2 Har Du haft några problem med att tvätta Dig och torka Dig (hela kroppen) på grund av Ditt knä?

Inga problem alls	Mycket lite problem	Måttliga problem	Mycket stora problem	Omöjligt att göra
☐	☐	☐	☐	☐

Under de senaste fyra veckorna...

3 Har Du haft något problem med att komma in i eller ut ur bil eller med att använda offentligt transportmedel (vilket Du nu tenderar att använda) på grund av Ditt knä?

Inga problem alls	Mycket lite problem	Måttliga problem	Mycket stora problem	Omöjligt att göra
☐	☐	☐	☐	☐

Under de senaste fyra veckorna...

4 Hur länge har Du kunnat promenera innan smärtan i Ditt knä blivit **svår**? (Med eller utan käpp)?

Ingen smärta/ >30 min	16 till 30 min	5 till 15 min	Endast runt huset	Inte alls - svår smärta direkt vid promenad
☐	☐	☐	☐	☐

ISBN 91-630-8224-1

Problem med ditt knä

Under de senaste fyra veckorna...

Markera en ruta för varje fråga

Under de senaste fyra veckorna...

5 Efter en måltid (sittande till bords), hur smärtsamt har det varit för Dig att resa Dig upp från stolen på grund av Ditt knä ?

Inte smärtsamt alls	Lätt smärtsamt	Måttligt smärtsamt	Väldigt smärtsamt	Outhärdligt
☐	☐	☐	☐	☐

Under de senaste fyra veckorna...

6 Har Du haltat då Du promenerat på grund av Ditt knä ?

Sällan / aldrig	Ibland eller endast i början	Ofta och inte bara i början	Merparten av tiden	Hela tiden
☐	☐	☐	☐	☐

Under de senaste fyra veckorna...

7 **Kan** Du sätta dig ner på huk och komma upp igen efteråt?

Ja, lätt	Med viss svårighet	Med måttlig svårighet	Med mycket stor svårighet	Nej, omöjligt
☐	☐	☐	☐	☐

Under de senaste fyra veckorna...

8 Har Du besvärats av smärta i Ditt knä då Du legat till sängs på natten?

Inga nätter	Bara 1 eller 2 nätter	Vissa nätter	De flesta nätter	Varje natt
☐	☐	☐	☐	☐

ISBN 91-630-8224

Problem med ditt knä

Under de senaste fyra veckorna...

Markera en ruta för varje fråga

Under de senaste fyra veckorna...

9 I vilken grad har smärtan i Ditt knä påverkat Ditt vanliga arbete (inklusive hushållsarbete)?

Inte alls	Lite grann	Måttligt	I hög grad	Fullständig
☐	☐	☐	☐	☐

Under de senaste fyra veckorna...

10 Har det känts som om Ditt knä plötsligt skulle "vika sig" eller svika Dig?

Sällan/ aldrig	Ibland eller bara i början	Ofta och inte bara i början	Merparten av tiden	Hela tiden
☐	☐	☐	☐	☐

Under de senaste fyra veckorna...

11 **Kan** Du handla det som behövs till hushållet på egen hand?

Ja, lätt	Med viss svårighet	Med måttlig svårighet	Med mycket stor svårighet	Nej, omöjligt
☐	☐	☐	☐	☐

Under de senaste fyra veckorna...

12 **Kan** Du gå nerför en trappa?

Ja, lätt	Med viss svårighet	Med måttlig svårighet	Med mycket stor svårighet	Nej, omöjligt
☐	☐	☐	☐	☐

ISBN 91-630-8224-1