From the Department of Orthopaedics, Institute of Surgical Sciences,
Göteborg University, Göteborg, Sweden

On the validity of the results from the Swedish National Total Hip Arthroplasty register

Peter Söderman

THESIS

ACTA ORTHOPAEDICA SCANDINAVICA SUPPLEMENTUM NO. 296, VOL. 71, 2000

Printed in Sweden
Wallin & Dalholm, Lund
2000

Contents

List of papers

The thesis is based on the following papers:

I **Validity and reliability of Swedish WOMAC osteoarthritis index. A self-administered disease-specific questionnaire (WOMAC) versus generic instruments (SF-36 and NHP).**
Söderman P, Malchau H.
Acta Orthop Scand 2000; 71 (1): 39-46.

II **Are the findings in the Swedish National Total Hip Arthroplasty Register valid? A comparison between the Swedish THA register, the national Discharge register and the national Death register.**
Söderman P, Malchau H, Herberts P, Johnell O.
J Arthroplasty 2000; 15 (7): 884-889.

III **Outcome after total hip arthroplasty. Part I. General health evaluation in relation to failure definition in the Swedish National Total Hip Arthroplasty register.**
Söderman P, Malchau H, Herberts P.
Acta Orthop Scand 2000; 71(4): 354-359.

IV **Outcome after total hip arthroplasty. Part II. Disease specific follow-up and the Swedish National Total Hip Arthroplasty register.**
Söderman P, Malchau H, Herberts P, Zügner R, Regnér H, Garellick G.
Accepted for publication in *Acta Orthop Scand.*

V **Outcome of total hip replacement. A comparison of different measurement methods.**
Söderman P, Malchau H, Herberts P.
Submitted to *Clin Orthop*

Abstract

The Swedish National Total Hip Arthroplasty Register contains more than 200,000 primary and secondary hip replacements. The failure end-point definition is revision.

The aim of this thesis was to validate the results presented by the register and to study the outcome of hip replacement surgery in Sweden. The hypothesis was firstly that the number of failure reported to the Swedish THA register are valid and secondly that adding clinical and radiographic failure criteria will dramatically decrease the survival rate for THR implants.

The study consisted of three parts with 2–10 years follow-up of patients with total hip replacements (THR). In part I, three general health questionnaires (Nottingham Health Profile, SF-36, EuroQol) and two disease-specific instruments (WOMAC, Harris Hip Score) were tested for validity and reliability (n=62). The results showed the disease specific questionnaires are at least as valid and reliable as the general instruments are.

In part II, all THRs reported to the Swedish THA register 1986 to 1994 (84,884 primary and 10,176 revision hip replacements) were compared with the data from the Discharge register and the Cause of Death register in Sweden. 2,604 patients were randomly selected from the Discharge register to determine if they had undergone any revision surgery. The study showed that the Swedish THA register covers 94% of the revisions actually performed in Sweden and the results did not differ significantly from the data in the Discharge register and the results reported by the patients.

In part III, 1,056 patients from the selected cohort were studied further concerning general health and disease-specific health, using the Nottingham Health Profile, SF-36 and WOMAC. An age and gender matched subcohort of 344 patients were then examined clinically, using the Harris Hip Score, and radiographically. The clinical and radiographic failure rates were in several tests as high as the revision rates documented in the Swedish THA register. The clinical results were, however, dependent on demographics, the definition of clinical failure and the scoring system used. The results presented by the register with revision as failure end-point give an exact but limited information about the quality of hip replacement surgery in Sweden.

Introduction

The methods for diagnosis and treatment of hip disability are continuously developing and total hip replacement is today one of the most effective surgical interventions ever introduced (Rissanen et al. 1995, Hozack et al. 1997, Garellick et al. 1998).

The results of total hip replacement (THR) surgery have been studied using different outcome tools focused on the process of care (pain, function, radiographic results) and on patient-oriented outcome (quality of life) and the need for revision of the implant. For research follow-up of new implants, a stepwise introduction has been proposed (Gross 1993, Malchau 1995). The steps are: laboratory studies, small cohort studies using radiostereometry (RSA), a very precise radiographic method, multicenter studies and national register studies. Register studies, like the Swedish National Total Hip Arthroplasty register, were introduced in Sweden twenty years ago (Ahnfelt 1986). The register provides hard data with revision as a limited but exact failure end-point definition. Although there has been a need for registers and evidence based medicine (Black 1996, Morris 1996, Sochart et al. 1996), the failure end-point has been criticized and the value of register data questioned (Bulstrode 1996).

There is an increasing frequency of hip surgery, more old people with degenerative hip disorders and more osteoporotic hip fractures, and a rapidly rising costs of health care (Lidgren 2000). A demand of simpler outcome instrument has become almost imperative (Keller et al. 1993). One type of such an instrument is a numerical or adjectival scale that is easy to administer and analyze in a standardized format. This type of soft data is useful for routine clinical follow-up to control the quality of hospital care.

The goal with THR must also be a satisfied patient. If one assume that the patient and the surgeon have the same goal, the treatment must lead to a satisfied patient. What is then a satisfactory outcome? The most common answer from the patients asked about what is making them satisfied is the relief of pain and good function (Amadio 1993, Rowley 1997, Mancuso et al. 1997).

In summary, outcome studies of THR should include soft data with information about pain and function. But it is also important, especially for continuous quality and research evaluation, with hard data such as the results given from national registers. The present study is unique because it provides information about hard and soft data after hip replacement on a national level. A brief review of the registers used in this study and outcome measurements in general is presented.

Registers used in the present study

The Swedish National Total Hip Arthroplasty register (the Swedish THA register)

Since Charnley introduced his total hip prosthesis (Charnley 1961), many different implants and surgical techniques have been developed, not only because of improved implants and techniques as such but also because of antiseptic, radiographic and anaesthetic improvements. It has been estimated that 1,000,000 patients annually undergo total hip replacement today and several hundred different implants are used. To detect complications after total hip replacement, the patients should be followed prospectively. But it would be very costly to follow each patient. Therefore, a *pilot study* was initiated in Sweden and performed during 19 months between 1976 and 1977 (Ahnfelt et al. 1980). The aim of the pilot study was to investigate the possibility of registering if a primary arthroplasty has failed or not and find out why this failure occurred. 5,758 primary operations and 577 reoperations were detected in 36 hospitals. The study gave important information about severe complications due to hip replacement surgery and the results motivated all hospitals in Sweden to participate in a nationwide prospective study on hip replacement surgery. The

Swedish National Total Hip Arthroplasty register was initiated in January 1979 by Herberts and Ahnfelt (Ahnfelt 1986, thesis). The register was early supported by the Swedish Orthopaedic Society and later by the National Board of Health and Welfare.

Today the aims of the register are:

- To perform epidemiological analyses of hip replacement surgery in Sweden.
- To identify risk factors for failure of primary and revision surgery.
- To facilitate improvement of the surgical technique by risk factor analysis.
- To perform bench marking by comparison between regions.
- To perform quality assurance of all hip replacements performed in Sweden.

Results from the register have been published at regular intervals (Herberts et al. 1989, Ahnfelt et al. 1990, Malchau et al. 1993, Herberts et al. 1997, Herberts and Malchau 2000). Further, the results from the register are presented annually, with compilations to the profession and the respective clinics, and there are continuous exhibitions at different scientific meetings such as the Annual Meeting of the American Academy of Orthopaedic Surgeons (Malchau et al. 2000).

Three different databases are the cornerstones of the register:

1. The number of **primary operations** from 1979 to 1991 were registered annually and per department, including the type of implant. Over 180,000 operations have been registered to date.

Since 1992, all primary operations have been reported in more detail, including the patient's id number and diagnosis. The unique identity number used for all permanent residents in Sweden gives information about age and gender. The implant have been characterised in detail during this later period. Since 1999, the clinics report to the register through the Internet and the register has its own website where the clinics can get feedback about their results. The public is thereby also informed about the results of hip replacement (http://www.jru.orthop.gu.se).

2. The number of secondary surgery (reoperations) is obtained, since January 1979, when all clinics in Sweden send copies of the medical records to the register. 132 variables were initially recorded concerning any **reoperation** after primary hip replacement. Logistic and economic reasons made radiographic follow-up impossible. More than 100 parameters per reoperation are today recorded from the medical records. Some essential parameters are now reported via the Internet site. Three specially trained secretaries collect the records and transform the information to one database. The unique id numbers have been registered for all reoperations and major revisions (exchange or removal of implant). Today, over 16,000 revisions are registered.

3. The third database concerns the prophylactic measures against aseptic loosening and infection annually. The variables included in the reports are: surgical approach, type of cement and cement mixing technique, use of brush and pulsative lavage, number and diameter of anchorage holes in the acetabulum, cement application technique including use of a distal femoral plug and proximal seal, type of antibiotic prophylaxis, length and administration mode. The variables are not reported for the individual patient but for the department as a whole and per year. This information is also available from the start in 1979.

Statistical methods used in the Swedish THA register

Patient-related factors and implant-related factors have been analysed by estimation of the survival function for all implants depending on age, gender, diagnosis, type of implant and fixation technique (Kaplan and Meier 1958). The register was compared with the expected survival of an age and gender matched Swedish population. The complication risk for different types of implants could therefore be calculated. Today, the effect of various surgical and cementing techniques on revision rates is analysed by Poisson models (Breslow and Day 1987). The hazard functions of revisions are thereby estimated by a stepwise procedure ending up with the significant variables in a multivariate model. The influence of these factors on the risk of revision for aseptic and septic loosening is calculated using multiple regression survival analysis.

Results of total hip replacement from the Swedish THA register 1979–1999

Demographics: During the last 10 years approxi-

Table 1. Number of primary THR with different fixation and age from 1992 to 1999

Age	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99	100-109	Total
Cemented	20	100	333	1,337	6,333	18,549	**30,723**	13,823	783	5	72,003
Uncemented	4	88	185	615	**1,132**	369	27	7	0	0	2,427
Hybrid	5	48	138	598	**1,699**	1,346	339	110	8	0	4,291
Total	29	236	656	2,550	9,161	20,264	31,089	791	5	78	78,721

mately 10,000 primary operations (100/100,000 inhabitants) and 1,000 revisions have been performed annually in Sweden. The mean age was 70 years and 60% were women. The three most common indications for THR in Sweden are osteoarthrosis, fracture and arthritis. The majority of older patients were given a cemented prosthesis while hybrids and uncemented implants are more common in the younger populations (Table 1).

Implant-related factors and results in the Swedish THA register: In Figure 1, all primary total hip replacement implants during three time periods are illustrated. The cut-off 1987 was chosen because at that time most orthopaedic units in Sweden adopted a modern cementing technique. The time period, 1992–1998, was chosen as modern uncemented technology, with cup designs and active surface coating on the stems, was introduced around 1992 (Thanner et al. 1999, Kärrholm et al. 1998, Nivbrant 1999). When the material was classified as cemented, uncemented and hybrid implants during the three time periods, a clear improvement in cemented prostheses was observed (Figure 1). The uncemented implants showed a minor improvement. The reason may be many different types of uncemented implants used during these years. Too many uncemented designs with inferior performance were still used in 1991, thus affecting the overall results. A similar finding was reported from the Norwegian register (Havelin et al 1993).

Almost all of the cemented implants show a significantly improved survival between the early period, 1979–1987, and the later period, 1988–1998. Using the modern cementing technique, a 10-year survival of 94.6% was obtained for hip replacement with the index diagnosis osteoarthrosis and revision due to aseptic loosening. The survival at 17 years varies between 80 and 87% for well-documented and common implants.

Cementless implants have been used to a limited extent, both during the first and in the second time period. Since the number of observations is limited, there is a broad confidence interval and more uncertain information. For certain uncemented implants, there is a statistically significant improvement with this separation in to three time periods. The third generation of uncemented implants used during the last seven years (often with hydroxyapatite coating or textured titanium sur-

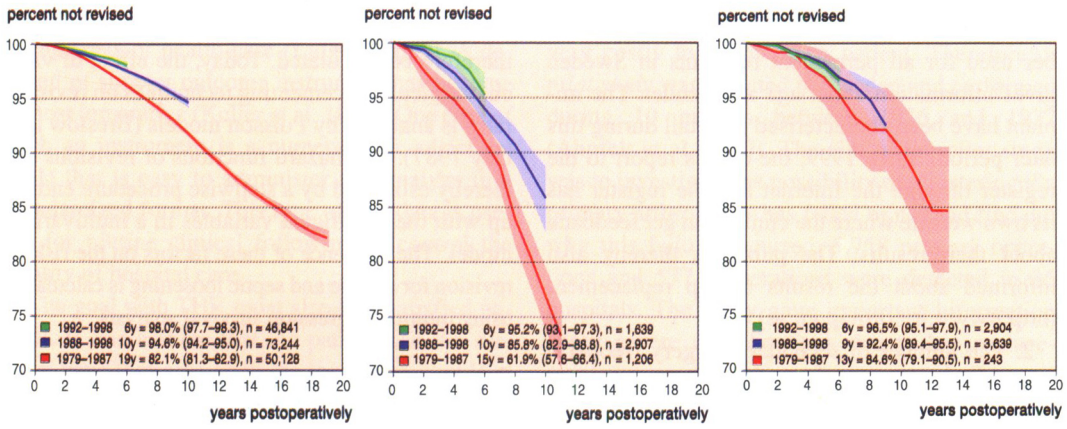

Figure 1. Survival of all primary THR separated in cemented (left), uncemented (middle) and hybrid (right).

faces) has functioned well in the short perspective up to five years (Thanner et al 1999).

In summary, most modern cemented implants have a ten-year survival rate for aseptic loosening around 94% and the best have shown a survival rate of approximately 98%. These figures illustrate the excellent results of hip replacement surgery in Sweden obtained by the average orthopaedic surgeon. The excellent short-term result for some cementless implants is promising, but the observation period is too short to accept uncemented fixation as a safe and efficacious procedure.

Patient-related factors and results in the Swedish THA register: Men are at significantly higher risk for revision procedures than women, when analysed for revisions attributable to aseptic loosening. During the later time period, this difference became less pronounced as a result of improved technique. With respect to age, the register shows that the younger and more active patients are at greater risk in all diagnostic groups. Approximately 12,000 patients are included in the cohort with a higher failure rate. It is above all for this cohort of patients that further research and development is mandatory. A possible solution could be referral of these patients to centres of excellence.

Environmental factors and results of THR in the Swedish THA register: In order to evaluate patient-related factors (gender, age and diagnosis) and fixation mode independently, a new Poisson model was recently applied to the cohort operated upon between 1992 and 1998 (Breslow 1987, Malchau et al. 2000). This cohort was chosen as both primary and revision data were reported by use of the specific patient id number. Certain general parameters (time since operation, calendar time) are a prerequisite for this specific statistical analysis. The following text summarises the results of the new Poisson models: the estimated risk of revision is 20% lower for women compared to men ($p<0.01$). This risk decreases with increasing age. This finding underlines the importance of taking considerations related to age, such as comorbidity and activity, when discussing different fixation modes (Havelin et al. 1993). The results of THR were also dependent on the type of cement used and vacuum mixing of the cement. The time dependent Poisson model indicates, as previously reported, a higher risk of revision in the first 4–5 years after operation. Further follow-up shows a continuous reduction of the risk of revision. At eight years the risk was 0.74 when vacuum was compared with manual mixing. Therefore, the use of vacuum mixing of cement seems to be justified.

Conclusions from the Swedish THA register: Joint replacement procedures lend themselves particularly well to national register studies. The following conclusions were reported (Herberts and Malchau 2000, Malchau et al. 2000).

- The number of THR procedures performed in Sweden has increased gradually and over time well-documented implants have been consistently used.
- The most serious complications (aseptic loosening and deep infection) have declined three-fold over the past two decades. Aseptic loosening constitutes the main problem.
- Specific patient cohorts have increased failure risks, especially younger patients.
- An improved surgical and cementing technique was adopted in Sweden partly as a result of this register effort. Outcome has improved and differences between units diminished.
- The revision rate over the whole study period is only 7 % for primary cemented implants, which sets the standard for this surgical procedure.

The Swedish Hospital Discharge register (HDR)

The Swedish Hospital Discharge register (from the Centre of Epidemiology at the Swedish National Board of Health and Welfare, www.sos.se/epc) was developed during the 80s. The register contains information about the patient's identity number, gender, age, county, hospital, ICD code for diagnosis and surgical interventions, date of admission and discharge, and from and to which type of care the patients are admitted or discharged respectively. Since 1986, it is mandatory for all hospitals in Sweden to report the above-mentioned variables to the register. To study the amount of not reported diagnosis (23%) and errors (14%) a validation of the register were performed in 1990 (www.sos.se/epc).

The Swedish Cause of Death register (CDR)

All deceased inhabitants are reported to the Swed-

ish Cause of Death register, whereby statistics in such fields as changes in the population structure are produced. General statistics are available to the public on the Internet (www.scb.se) and after obtaining permission different registers can compare themselves with the CDR to study date and cause of death.

Outcome measurements

Although the surgical technique and implants are today almost optimal we still have failures. There has been growing interest in outcome studies after hip replacement and a demand for evidence-based health care. There are two main types of outcome instruments or questionnaires:

1. *Disease-specific questionnaires* with questions about the studied disease without influence from other diseases and with higher responsiveness than general instruments. Taylor's pain index (1937) and Keele's 5-point functional scale (1948) are two examples of early disease-specific scoring systems (Bellamy 1993). Since 1969, the staff-administered Harris Hip Score system is probably the most widely used scoring system for evaluating results of hip replacement (Harris 1969). Today, self-administered scoring systems like the WOMAC (Bellamy 1988) have become more popular. The Western Ontario and McMaster University Osteoarthritis Index (WOMAC) is a disease-specific, self-administered, health measure developed to study patients with osteoarthritis in the hip or knee. The original version contained 41 items in 5 domains. The domains were pain, stiffness, physical function, social function and emotional function. Internal consistency (Cronbach's alpha) and test-retest reliability (Kendall's tau c statistic) were moderate to excellent for pain (0.86 and 0.68, respectively), stiffness (0.90 and 0.48), physical function (0.95 and 0.68), social function (0.89 and 0.61) and emotional function (0.96 and 0.72). The Doyle Index, Lequesne Index and Bradburn Index have been used to demonstrate statistically significant correlations with WOMAC items in the same dimensions (convergent construct validity) and other dimensions (divergent construct validity). Construct validity was high for the domains of pain, stiffness and physical function. Social and emotional domains had lower validity. Hence the final index utilised the pain (5 questions), stiffness (2 questions) and physical function (17 questions) domains (Bellamy et al. 1988, Sun et al. 1997).

2. *General or generic questionnaires* asking patients about their health-related quality of life (HRQL), were the results can be compared with other conditions. The early general outcome questionnaires were long (approximately 100 items), such as the Sickness Impact Profile (SIP) and Medical Outcome Study (MOS), but in recent years there has been a tendency to use shorter tests. Two tests that are short are the Medical Outcomes Study 36-Item Short-Form Health Survey (SF-36) and the Nottingham Health Profile (NHP) with 36 and 45 questions respectively.

The SF-36 is a general self-administered questionnaire developed for applications in psychometric theory. It was developed from the Rand Corporation's health insurance experiment. The original measure was lengthy, containing 108 items. Later on, the test was used to studying effects which seemed to be a direct function of disease and treatment, that is health-related quality of life (HRQL) (Ware and Sherbourne 1992, Brazier et al.1992).

The NHP is a self-administered general instrument used to study quality of life after medical and/or surgical treatments such as total hip replacement (Wiklund and Romanus 1988). Like the SF-36 and WOMAC, the NHP has high validity and reliability (Hunt et al. 1980, Hunt et al. 1981).

The EuroQol (EQ or EQ-5D) is a generic multidimensional HRQL profile. The items were chosen from a review of the QWB, SIP, NHP and Rosser Index, and a single score is generated for each health state (Andersson et al. 1993). The EuroQol is a very short scoring system (5 items) which has been criticised for being less responsive than short questionnaires such as the SF-36, especially at the ceiling, but it has better results concerning the floor effects. Acceptable values for construct validity and internal consistency reliability have been reported, although they are lower than those of the SF-36 and NHP (Brazier et al. 1993, Brazier et al. 1996, Chetter et al. 1997, Essink-Bot et al. 1997). However, the brevity and

simplicity of the EuroQol questionnaire help it to achieve a better response in stroke survivors than with the SF-36. Some reports on the EuroQol state that it is valid, responsive to change and sufficiently reliable for group comparisons (Dorman et al. 1997, Hurst et al. 1997, Krabbe et al. 1996, Uyl-de Groot et al. 1994, VanAgt et al. 1994).

As in all clinical research, demographic bias should be taken under consideration and the scoring system used should have high validity, reliability and responsiveness. Although Chronbach (1951) emphasised the importance of a correctly produced questionnaire, the early scoring systems were not validated. The relevant expressions are defined as follows:

Outcome is the sum of the observations of an individual associated with a study period; a measure of change, related to a baseline measurement, and to specified treatment goals (Öberg 1996).

Demographic information is compiled by collecting health profile information including age, gender, diagnosis, current health conditions and comorbidities (Ritter and Albohm 1997).

Validity

A questionnaire with high validity is a questionnaire that measures what it is supposed to (Guyatt et al. 1989). For example: if one wishes to study pain after hip replacement, one should use questionnaires asking the patient if he or she has any pain. These questionnaires could be formulated in different ways. One questionnaire could ask the patient about pain in general, that is not only pain from the affected hip. Another questionnaire could ask about pain only in the affected hip. If the results are equal, the questionnaires are valid.

The most difficult aspect of validity is the terminology. Until the 1970s, almost all textbooks divided this topic into content validity, criterion validity and construct validity (Landy 1986). More recently, new "types" of validity have been proposed such as dividing construct validity into discriminant and convergent validity. However, it is important to point out that validation of a scoring system is a process of hypothesis testing where the system is tested in different ways (Streiner and Norman 1989). Some common types of validity are defined below.

Face validity is also termed clinical or biological validity (Bellamy 1993). This type of validity is based on the individual experience of a scoring system.

Content validity is a non-statistical comparison between the content in two scales. Do the scales concern the same area of interests (same domains or subgroups)? Questionnaires with good content validity are expected to have fewer categories with ceiling or floor effects. A floor effect occurs when the patient reports the poorest function for all or almost all items and receives the worst possible score, making it impossible to demonstrate deterioration in function over time for such a patient with use of the questionnaire. A ceiling effect occurs when the patient reports excellent function and receives the best score, making it impossible to show improvement in function over time (Martin et al. 1997).

Construct validity is concerned with the extent to which a particular measure relates to other measures with theoretically derived hypotheses concerning the concepts (or constructs) (Carmines and Zeller 1979). There are two types of validity often discussed as construct validity: *Convergent validity* is shown if two domains in different scoring systems have high correlation, that is if they show the same thing, e.g. pain. *Divergent (or discriminant) validity* is demonstrated if the correlation between scores on the same health component as measured by two different instruments (e.g. pain compared to pain) is higher than between scores on that health component and each of several other components (e.g. pain compared to social function).

Criterion validity can be demonstrated if a questionnaire shows high correlation with a gold standard. Two problems with this type of validity are how to set the gold standard and which level a correlation should reach for statistical significance. Bellamy has proposed a correlation over 0.80 since one is dealing with a gold standard (Bellamy 1993).

Reliability

Reliability is an expression of the extent to which the same results are yielded on repeated application of an assessment technique assuming no true interval change in the phenomenon under study

(Bellamy 1993). That is, does the instrument show the same results between different times of investigation and between different investigators if the patients' statuses are stable. Three types of reliability are defined:

Test-retest reliability (reproducibility, intra-observer reliability) is shown if the results are equal between two different times of investigation. The interval between the two examinations should not be too short because the patient might remember the questions. However, a long interval increases the risk of changes in patients' health. A commonly used interval is therefore 3–4 weeks.

Inter-observer reliability is the correlation between different investigators' examinations of the same patient.

Internal consistency reliability (Split-half) avoids the time-dependent changes presented with test-retest reliability. Internal consistency reliability or intraclass correlation describes the correlation between items in the same domain. Chronbach's alpha index is often used for this purpose. That is, a domain of interest should contain more than one item to reduce the risk that the patient had misunderstood one item. Several questions about the same thing (i.e. pain) can act as control items each other (internal consistency).

Responsiveness

Responsiveness means how well a scoring system reacts to clinical changes, that is the sensitivity to change of an assessment technique.

The correlation calculations in validity and reliability tests could be done using different coefficients as long as the author describes which coefficients were used. Two of the most common coefficients are Pearson's r and Spearman's rank order correlation coefficient, rho (Katz et al. 1996). Spearman's rho should be used instead of Pearson's r on an ordinal scale and if the correlation is not linear.

How important is the level of correlation? With the formula $n=(1/r)^2$, one can calculate the number of patients that are needed in a study in the worst case scenario to compensate for differences between scoring systems when the correlation coefficient (r) is known. If r is 0.5, one should in the worst case scenario use 4 times, $= (1/0.5)^2$, more questionnaires to ensure that different results are not due to a bad scoring system. This is an example of the importance of using a correct evaluation tool for follow-up of THR. Hence, the first step in the study leading to this thesis was to determine which scoring system should be used in the clinical evaluation of the Swedish THA register.

Aim of the study

The end-point definition of failure in the Swedish National Total Hip Arthroplasty Register is revision. The overall aim of this thesis was to validate the results presented by the register by comparing the clinical and radiographic outcome with the results presented by the register. Does the register capture all revisions in Sweden? Is revision a reliable end-point? Will the survival rate change significantly if clinical and/or radiographic failure end-points are used?

The hypothesis for this thesis was firstly that the number of failures reported to the Swedish THA register are valid and secondly that adding clinical and radiographic failure criteria will dramatically decrease the survival rate for THR implants.

The specific aims were:

1. To determine the number of primary THR not being revised and furthermore find those not reported to the Swedish National Total Hip Arthroplasty register, that are in need of future revision.
   - How many patients subjected to THR experience clinical failure?
   - How many patients experience radiographic failure?
2. To study the outcome of THR in a national study not based on university hospitals.
   - Are there any differences between regional, county, and rural hospitals?
   - What is the outcome for patients evaluated by self and staff administered assessments?
   - Are the mortality rates for primary and revision THR increased compared to an age and gender adjusted sample from the general population?
3. To develop an algorithm for follow-up of THR patients.
   - Can the physician be replaced by a physiotherapist using a staff-administered scoring system?
   - Is it possible to replace the staff-administered follow-up examinations with a self-administered questionnaire?
   - Is there any correlation between radiographic failure and clinical failure? Can clinical follow-up replace radiographic follow-up?
   - To define a suitable scoring system for patients with THR by testing validity, reliability and responsiveness.

Methods and patients

After review of the literature and discussion with the involved orthopaedic surgeons from the Swedish Knee Arthroplasty Register (L Lidgren, K Knutson, L Ryd, O Robertsson), the originators of the Swedish SF-36 (M Sullivan) and NHP (I Wiklund) and statistical advisers (A Odén among others), the study was presented at the Linköping meeting on outcome measurement in 1997 (Söderman and Malchau 1998).

The number of patients required was estimated by power analysis (A Odén) and the study was planned to consist of three parts (Figure 2). The first part includes Paper I, the second part Paper II and the third part consists of Paper III–V.

Paper I (validation of scales)

The following health-related quality of life (HRQL) assessment instruments and disease-specific instruments were used and tested for validity and reliability to ensure their suitability for part three of the study.

Methods – Outcome measurement scales

The SF-36 consists of 36 questions divided into eight domains: physical function, social function, role-emotional, role-physical, bodily pain, general health, mental health and vitality. The SF-36 is the most frequently used health-status measure in Northern America and has been used after total

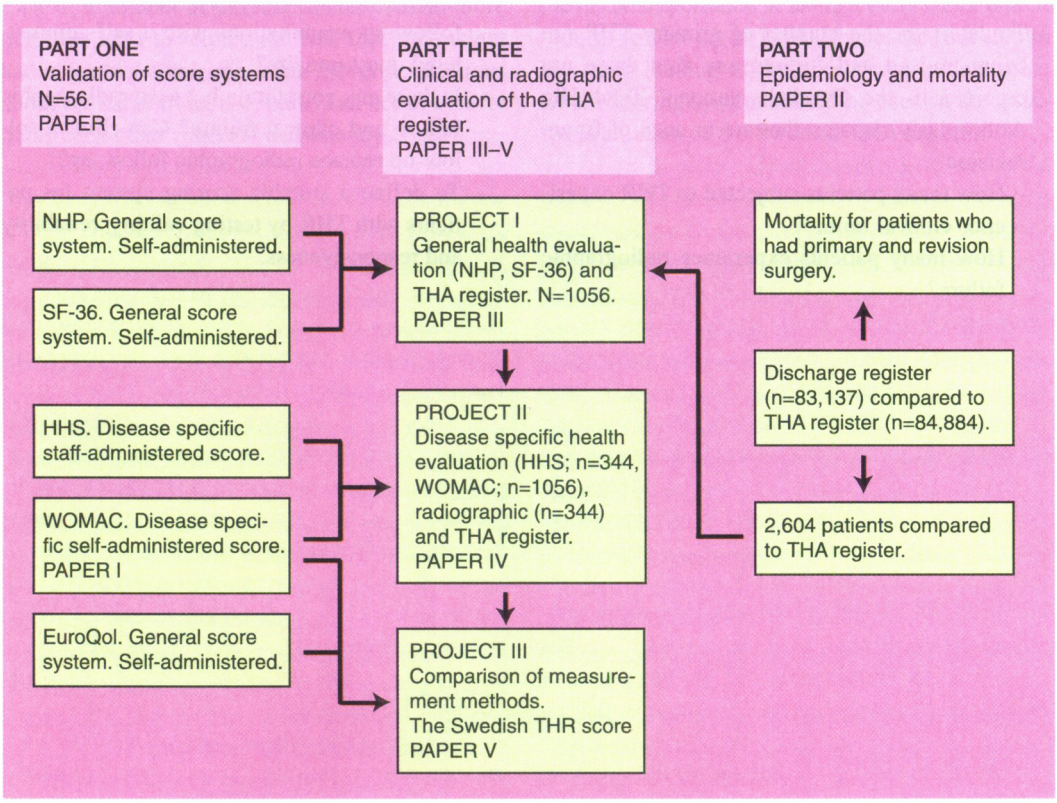

Figure 2. Patients and organization of the study – general overview.

joint replacement in the hip and knee (Martin et al. 1997, Ritter et al. 1995, Ware and Serbourne 1992). Sullivan et al. (1995) translated the test into Swedish and tested its validity and reliability. The raw score was transformed to a 0-100 scale (transformed scale score) as recommended by the Swedish manual (Sullivan 1994). A high value indicated a better result.

The Nottingham Health Profile (NHP) is a self-administered general instrument. The test consists of two parts with 45 yes-or-no answers. There are 38 questions in part one concerning patients difficulty in life and it is divided into 6 domains: emotional reaction, sleep, energy, pain, physical mobility and social isolation. The items are weighted and each dimension yields a value between 0 and 100. Part two consists of seven statements that reflect the frequency of problems in various areas of life: occupation, housework, social life, family life, sexual function, hobbies and holidays (McKenna et al.1981). To avoid negative values in the comparison with the WOMAC and the SF-36, the score was inverted so that maximum health yielded 100 points.

The EuroQol consists of five non-disease-specific dimensions similar to those in conventional generic health measures like the NHP: mobility (m, question 1), self-care (so, question 2), usual activities (ua, question 3), pain/discomfort (p, question 4) and mood (a, question 5). Each dimension has three categories of the general form: level 1 = no problem, level 2 = some problem, level 3 = extreme problems. Altogether 3^5 (243) descriptions are possible. Data from the EQ-5D can be presented in three forms: 1) the patient's level of the problem in each of the five domains (EQ-5D profile), 2) as a health index (EQ-5D utility), and 3) patient's evaluation of their own global health status (EQ-5Dvas) (Hurst et al. 1997, Krabbe et al. 1996, VanAgt et al. 1994). In this study, the EuroQol gives a minimum of 0 and a maximum of 100 points for the total score in each domain.

The Western Ontario and McMaster University Osteoarthritis Index (WOMAC) is a self-administered, disease-specific instrument. There is one computerised version of the WOMAC and four basic versions with different types of scales on which the response is scored (LK = 5-point Likert scale, VA = 10 cm horizontal visual analogue scale) (Bellamy et al. 1992, Bellamy et al. 1997). In the present study, the Likert scaled format was used. This questionnaire is reliable, valid and sensitive to clinically important changes in health status after surgical interventions (responsiveness) (McGrory and Harris 1996, Bellamy et al. 1988, Bellamy et al. 1991, Martin et al. 1997). The WOMAC was translated into Swedish according to the procedure presented by Guillemin (Roos et al. 2000). The three domains in the WOMAC can be analysed separately or according to a single score. Every question allows five alternative answers, which give a total of 0–4 points. The maximum score in the Likert version is 20 points for pain, 8 for stiffness and 68 points for physical function. In this study, the score, in each domain, was transformed to a 0–100-point scale. To make the results comparable with the SF-36 and NHP, the score was inverted. Therefore, a maximum score of 100 points occurred when the patient had minimum pain and stiffness and optimal function.

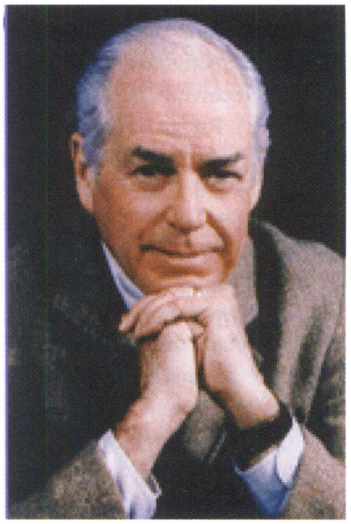

Figure 3. William Harris. Picture published with permission from W. Harris.

The Harris Hip Score (HHS) is a disease-specific test introduced by W Harris (Figure 3) in order to provide an evaluation system for various hip disabilities and methods of treatment (Andersson 1972, Harris 1969). This rating system is not self-administered (staff-administered). The Harris Hip Score gives a maximum of 100 points and the

domains include pain, function, deformity and motion. Pain and function were the two basic considerations and received the heaviest weighting (44 and 47 points). Range of motion and deformity are seldom of primary importance, and hence received 5 and 4 points respectively. Function was subdivided into activity of daily living (ADL, 14 points) and gait (33 points). A total Harris Hip Score below 70 points was considered a poor result, 70 to 80 fair, 80 to 90 good and 90 to 100 excellent.

Patients

62 patients were randomly selected from Sahlgrenska University hospital for inclusion in the study. 4 patients were excluded because of intellectual deficiencies. 18 patients were previously subjected to total hip arthroplasty 10 years ago, using Charnley (DePuy Olmed), Lubinus IP (Link), Christiansen (Howmedica) or Exeter polished (Stryker Howmedica) prostheses. 40 patients were operated upon two years earlier with the Spectron prosthesis (Smith & Nephew). All patients filled in the SF-36, NHP, EuroQol and WOMAC during one week and they were also examined by two physiotherapists and two orthopaedic surgeons using the HHS. The procedure was repeated after four weeks. The mean age for the 58 patients was 71 (52–86) years. 38 were women. The 2-year and 10-year groups had the same gender distribution and the mean age was 73 and 70, respectively. One third of the total group was affected in one hip (Charnley category A) and one-third in both hips (Charnley category B). The remaining third had general disease or another disease that impaired gait (Charnley category C) (Charnley 1979). Power analysis predicted that, with 35 patients, there should be at least a 70 % chance of detecting a correlation between health status instrument score differences if one existed, assuming a Pearson's r value of 0.40 (McGrory and Harris 1996).

Paper II (epidemiology and mortality)

The epidemiology and mortality of THR were evaluated by comparing the Swedish National Total Hip Arthroplasty register with a database from the Swedish Hospital Discharge register and the Swedish National Cause of Death register.

Methods – Mortality

From the Discharge register, the expected and observed death rates in all hospitals (1986–1994) were calculated and standardised for age, gender, year of surgery and index diagnosis. Risk-ratios (observed/expected) and 95% confidence intervals were indicated. Regression analysis (Poisson models) was used to estimate the death risk in relation to patient age and length of follow-up.

Register validation

The national cohort for patients admitted to the hospital with a hip or knee diagnosis between 1979 and 1995 was obtained from the Discharge register (n = 890,000). From this cohort, all patients undergoing primary or revision hip replacement and re-operation between 1986 and 1995 were selected. These two subcohorts were compared with the data in the Swedish National Total Hip Arthroplasty Register during the same period. Gender, age, number of registered operations (primary and revision) and the annual procedure incidence (operations per 100,000 inhabitants) were calculated.

Two steps were taken to determine how many operations were missing in the hip register and the reason. In the first step, 61 medical records for patients registered in the Discharge register but not in the Swedish THA register were studied in detail to determine the reasons why the patients were not registered.

Secondly, from the Discharge register a randomly selected cohort of 2,604 patients subjected to THR between 1986 and 1995 received a short questionnaire asking, among other things, if they had undergone re-operation. The medical records for these patients were collected and studied, thus providing information about the type of repeat surgery that had been performed. Several types of major or minor reoperations can be done but in this study we were only interested in the most important, revision of the implant. With logistic regression analysis, the ten-year survival of the hip replacement surgery reported to the Discharge register was compared with the Swedish THA register.

Papers III–V (clinical and radiographic analyses)

The outcome of THR was compared with the Swedish THA register in three projects:

Project I: General health evaluation (Paper III)

The aim of this project was to validate the end-point for failure in the Swedish THA register and to measure the precision of the failure end-point by measuring the general health in non-revised patients after total hip arthroplasty performed in daily practice.

Patients: 1,056 patients subjected to primary hip prosthesis in Sweden between 1986 and 1995 were randomly selected from the Discharge register (the Swedish National Board of Health and Welfare) (Söderman and Malchau 1998). The selected patients were compared with the Swedish Cause of Death register to include only living patients. The patients received a short questionnaire concerning reoperation, pain and overall satisfaction. The patients also received two self-administered general health questionnaires, the Nottingham Health Profile (NHP) and the SF-36. Two more letters were mailed to the patients who did not answer within 3 weeks. Patients who still did not reply were phoned up and asked to send in the questionnaire.

Project II: Disease-specific and radiographic outcome of THR (Paper IV)

The aim of this project was to perform an outcome analysis using disease-specific analyses and radiographic evaluation, and thereby assess the clinical relevance of the strict end-point for failure in the Swedish THA register.

Patients: The same patients as in project I (n=1,056) were asked to fill in a disease-specific self-administered questionnaire (WOMAC). Two letters of reminder were sent to those who did not answer within three weeks. Patients who still did not reply were phoned personally and asked to send in the questionnaire.

An age and gender matched subcohort of 344 patients from the original group were randomly selected from nine cities which included regional, county and rural hospitals. One independent physician (PS) or one independent physiotherapist (RZ) examined this cohort clinically using the Harris Hip Score system. The patients were also examined using conventional radiographic techniques (n=344).

Radiographic method: Standard anteroposterior frontal, pelvic (centered over the symphysis) and true lateral radiography were performed. The Hodgkinson criteria for loosening of the cup were used (Hodgkinson et al. 1988). Postoperative radiographs were not saved in all hospitals and hence migration of the cup could not be measured as the patients were examined only once, 2–10 years postoperatively. Failure of the cup was classified as Hodgkinson type 3 with a 100% circumferential radiolucent line. The criteria for stem failure were debounding, stem fracture, cement fracture or a 100% circumferential radiolucent line (Garellick et al. 1999, Mullroy and Harris 1997). The patients were separated into two groups: one with radiographic failure and one without radiographic failure. The radiographic results were compared with the pain, function and total scores from the clinical investigations (Harris Hip Score and WOMAC). The frequency of revision was noted for each hospital type.

Project III: Comparison of different measurement methods (Paper V)

To develop recommendations for optimal follow up of total hip replacement projects I and II were compared with each other and with the results in part two. The investigations were performed within 3–4 months in order to minimise the risk of changes in the patients' status and allow comparison between the results.

Statistical analyses

The statistical analyses were done by consulting professional statisticians and by using SPSS (Statistical Package for the Social Sciences) for Windows (Chicago).

Validity and reliability

Content validity was tested by directly comparing the HHS, NHP, WOMAC and SF-36, and by studying the floor effects, ceiling effects, mean, median and standard deviations in each domain.

Construct validity can be evaluated by correlating the questionnaire scores with the characteristics of the patients (number of comorbid conditions, perceived overall health status, and changes in activities). Patients who have more comorbid conditions, poorer perceived over all health status, or changes in activities are expected to receive poorer scores. This is especially true for general questionnaires. Pearson's correlation between the total score of the tested instrument, such as the WOMAC, and the domain of interest in other questionnaires, such as function in the NHP and SF-36, should be significant at the level of 0.01 (Martin et al. 1997).

Total scores for the HHS, NHP, WOMAC and SF-36 were analysed in all patients and for male, female, two- and ten-years postoperatively, age over and less than seventy years. The patients were classified according to the Charnley category, that is (A) one hip affected, (B) both hips affected, and (C) multiple joint disease or other disabilities leading to difficulties in ambulating. Differences between the mentioned groups were calculated and significance analysis was done with the Mann-Whitney U-test. Divergent construct validity and convergent construct validity in domains of pain and physical function, and total score were tested using Pearson's and Spearman's correlation coefficients. The hypothesis was that the same domains should be more strongly correlated to each other, for example pain in the WOMAC, SF-36 and NHP, than with other domains such as function (Katz et al. 1996).

Criterion validity is present when the scores correlate with an accepted measure (gold standard) of the condition being evaluated. For criterion validity, an acceptable level for Spearman's rho is greater than 0.40 and $p<0.001$. SF-36 has been extensively validated and is a commonly used scoring system, for example after total hip arthroplasty. It was used as the gold standard in this study (Martin et al. 1997, Chetter et al. 1997).

To study test-retest reliability, 58 patients were examined with the three questionnaires, twice within four weeks. The patients answered the items in the WOMAC and SF-36 the same week that a physician and a physiotherapist examined them with the Harris Hip Score. Total score, domains and items were calculated with Pearson's and Spearman's correlation coefficients.

Internal consistency reliability was tested for the questionnaires and within the different domains in the HHS, WOMAC, SF-36 and NHP. Cronbach's alpha coefficient was used for this purpose.

Inter-observer reliability between physicians and physiotherapists was tested. The patients were divided into two groups, where one physician and one physiotherapist examining each group. The procedure was repeated after four weeks. Goodman-Kruskal's gamma, Pearson's and Spearman's correlations for items, domains and total score of the Harris Hip Score were calculated, and the inter-observer reliability regarding Charnley classification, bone length and the Trendelenburg test was measured.

Pearson's correlation coefficient was used to compare domains. The following guidelines were used to interpret the correlation coefficients (r): poor correlation ($r<0.3$), moderate correlation ($0.3<r<0.6$), good correlation ($0.6<r<0.8$), excellent correlation ($r>0.8$) (Bellamy et al. 1991).

Clinical outcome and survival analyses

Total score, domain scores, mean, median, standard deviation (95% CI), minimum value and maximum value (range) were calculated for all patients two to ten years postoperatively. The Mann-Whitney U test was used for statistical analyses (non-parametric tests). For the 10-year survival analyses, logistic regression analyses were used and the results were compared with the survival statistics in the Swedish THA register. Patients that were revised or scored under 50, 60 or 70 points were counted as failures. The patients with clinical failure according to this definition and patients with no failure were analysed with regression analysis, which provided a constant (B0), and B1 (follow-up time) and B2 (follow-up time x follow-up time). By using different time (0.25–10 years) and these beta values, the probability of survival (S) was estimated with the formula:

$$S = (1/(1+EXP(-(B0+B1 \times time + B2 \times time \times time))) \times 100$$

To make the results comparable with other scoring systems all domain and total scores were transformed to 0–100-point scales where 100

points indicated best health. The formula for transformation was:

$$\text{Transformed scale} = 100 \times \frac{(\text{actual raw score} - \text{lowest possible raw score})}{\text{possible raw score range}}$$

Overall satisfaction

The correlation between overall satisfaction (patient satisfied with the operation or not) and the domains in the WOMAC, NHP, Harris Hip Score and SF-36 were estimated using Spearman's rho.

Scoring system for a new questionnaire, the Total Hip Replacement score

The three pain and function items in WOMAC that had the highest reliability, validity, responsiveness and response rate were identified. Based on this information, a new scoring system for total hip arthroplasty was suggested. The method of development was as follows:

a) the test-retest reliability for each item was examined by determine the correlation between the answers to the WOMAC questionnaire administrated twice 4 weeks apart, n=56.
b) the content validity for each item was tested for the number of ceiling and floor effects, n=1,056.
c) the response rate was determined for each item, n=1,056.
d) the responsiveness was tested by studying the differences between Charnley category A and C, different ages, different follow-up times (2 and 10 years) and SF-36 above and below 70 points. Mann-Whitney's U-test was used, n=1,056.

Results

Paper I: Validity and reliability of Swedish WOMAC osteoarthritis index. A self-administered disease-specific questionnaire (WOMAC) versus generic instruments (SF-36 and NHP)

The aim was to determinate whether a disease-specific, self-administered questionnaire could replace a general instrument as an outcome measure after total hip replacement, and to test the validity and reliability of the Swedish WOMAC osteoarthritis index.

All 58 patients answered the full questionnaire. A few patients did not answer some of the individual items.

Validity

Content validity. 1 floor value was seen in the WOMAC, 2 in the SF-36 and 6 in the NHP. There were fewer ceiling values for function in the WOMAC and SF-36 than in the NHP, while the pain domain scored equal. The median score in the domains energy, social isolation and emotional reactions in the NHP and social function in the SF-36 was 100 points.

Construct validity. The SF-36 and NHP scored a greater difference between gender and the Charnley categories than the disease-specific scoring system, WOMAC. There were no significant differences between the three questionnaires with respect to age and follow-up time. The pain domain in the WOMAC correlated better with pain in the SF-36 (Pearson's r=0.62, Spearman's rho=0.59) and NHP (r=0.71, rho=0.62) than with function in the SF-36 (r=0.56, rho=0.45) and NHP (r=0.59, rho=0.55) (Table 2). The same results were obtained when comparing the function domain in the WOMAC and NHP scores with pain domains, but not between the WOMAC and SF-36. The correlation between domains in one questionnaire and its total score was higher than with the other questionnaires. For example, there was better correlation between the WOMAC score for pain and the total WOMAC score than between the WOMAC score for pain and the total NHP score. This study showed high convergent and divergent validity for the disease-specific test.

Criterion validity. The Spearman´s rho was acceptable when the total scores for the SF-36 and WOMAC (0.73) or the NHP (0.80) were correlated (Table 2). This was also true for the correlation between SF-36 domains and the same domains in the NHP and WOMAC (Spearman´s rho range 0.59-0.82).

Reliability

Test-retest reliability. Total score, domains and items were each tested with Pearson's and Spearman's correlation coefficients between two exam-

Table 2. Correlation between domains and total score in SF-36, NHP and WOMAC measured with Spearman's correlation

Score	domain	SF-36 pain	NHP	WOMAC	SF-36 function	NHP	WOMAC	SF-36 total	NHP	WOMAC
SF-36	pain	1								
NHP		0.69	1							
WOMAC		0.59	0.62	1						
SF-36	function	0.57	0.54	0.45	1					
NHP		0.56	0.59	0.55	0.73	1				
WOMAC		0.71	0.71	0.76	0.67	0.82	1			
SF-36	total	0.81	0.66	0.58	0.75	0.76	0.83	1		
NHP		0.7	0.79	0.66	0.59	0.74	0.78	0.8	1	
WOMAC		0.66	0.67	0.83	0.59	0.7	0.93	0.73	0.71	1

Table 3. Test-retest reliability for WOMAC, SF-36 and NHP. Pearson's (r) and Spearman's (rho) correlation between domains and total score

Test		WOMAC		SF-36		NHP	
		r	rho	r	rho	r	rho
	total	0.90	0.87	0.92	0.91	0.94	0.88
WOMAC	pain	0.88	0.78				
	stiffness	0.76	0.69				
	function	0.91	0.91				
SF-36	physical function			0.78	0.78		
	role-function			0.74	0.72		
	bodily pain			0.85	0.83		
	general health			0.84	0.81		
	vitality			0.92	0.92		
	social function			0.70	0.65		
	role-emotional			0.49	0.40		
	mental health			0.80	0.69		
NHP	emotional reaction					0.9	0.81
	sleep					0.82	0.86
	energy					0.78	0.87
	pain					0.89	0.83
	physical motion					0.88	0.83
	social isolation					0.82	0.80

inations, 3 to 4 weeks apart. The reliability for total score was excellent in every questionnaire (WOMAC r=0.90, SF-36 r=0.92, and NHP r=0.94). The reliability in domains was also high, and the disease-specific test had the highest value (Table 3).

The test-retest reliability for each item was moderate to excellent (Pearson's r range 0.47–0.81). The correlations were significant at the 0.01 level (2-tailed).

Internal consistency reliability. Cronbach's alpha coefficient showed high correlation between each domain (0.60–0.96). The WOMAC items received higher values (0.88–0.96) than the items in the NHP (0.65–0.83) and the SF-36 (0.75–0.94). The WOMAC scored higher reliability for pain and function than the SF-36 and NHP.

Conclusions. The Swedish WOMAC osteoarthritis index was highly capable of measuring what it was supposed to measure (high validity) and it was also reproducible (high reliability). Based on earlier investigations as well as the present study, we recommend the Swedish WOMAC Likert score for future studies after total hip arthroplasty. Information about the WOMAC questionnaire can be downloaded on the internet site www.QLMed.org/WOMAC.

Paper II: Are the findings in the Swedish National Total Hip Arthroplasty register valid? A comparison between the Swedish THA register, the national Discharge register and the national Death register

The aim was to validate of the Swedish THA register by comparison with the Discharge register studying the epidemiology and mortality after hip replacement.

Primary operations

A total of 84,884 and 83,137 primary operations were registered from 1986 to 1994 in the Swedish THA register and the Discharge register, respectively. The primary procedure incidence for hip replacement in different regions in Sweden was found to vary between 81 and 129 per 100,000 inhabitants per year. Figure 4 shows the differences between the three major cities in Sweden (Stockholm, Göteborg and Malmö) compared to the average for the whole country. The incidence showed a slight increase but it was much lower in these cities than throughout the country in general.

Revisions

In 1996, a total of 10,176 and 11,323 revision op-

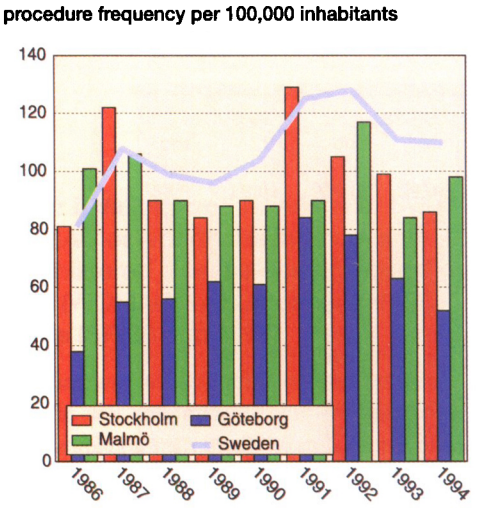

Figure 4. Procedure frequency (per 100,000 inhabitants) for primary THR in the Swedish THA register 1986–1994.

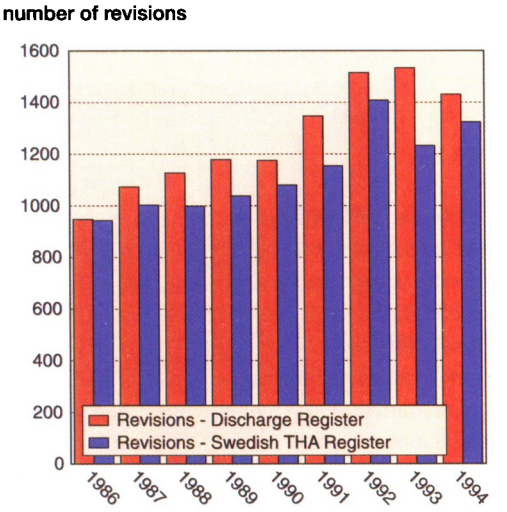

Figure 5. Number of hip revisions (extraction or exchange of prosthesis) in the Discharge register and the Swedish THA register between 1986–1994.

erations were registered in the Swedish THA register and the Discharge register, respectively. The number increased to 1991–1992 (Figure 5). The Swedish THA register showed a lower number of revisions, which could be an effect of different definitions of revision. 96% of the 2,604 patients reported to the Discharge register answered the questionnaire about revision or not. The results showed 10% missing revisions during 1986 and 1995. Of the 42 hospitals randomly selected for the self-administered questionnaire, 2 were responsible for 46% of the non-reported revisions. These missing revisions are now included in the Swedish THA register. This means that 95 % of the units that reported currently to the Swedish THA register during these years accounted on average for 94% of the revisions performed.

The ten-year survival based on the Discharge register cohort (n = 2,604) was not significantly different from that according to the Swedish National Total Hip Arthroplasty Register (n=93,852) (Figure 6). There were no significant differences in survival rates based on the total cohorts from the Swedish THA register (n=93,852) and the Discharge register (n=87,129). The survival according to this study of the two registers was 91–94% after ten years. Even if there were no significant different survival between the two registers, the Discharge register showed better survival because of two different survival analyses were used (the risk for death were not included in the logistic regression analysis in the analysis of the Discharge register).

Mortality

The overall risk of death after primary hip replacement was found to be 1% higher for men and 6% higher for women when compared with an age and sex-matched cohort (Table 4). The corre-

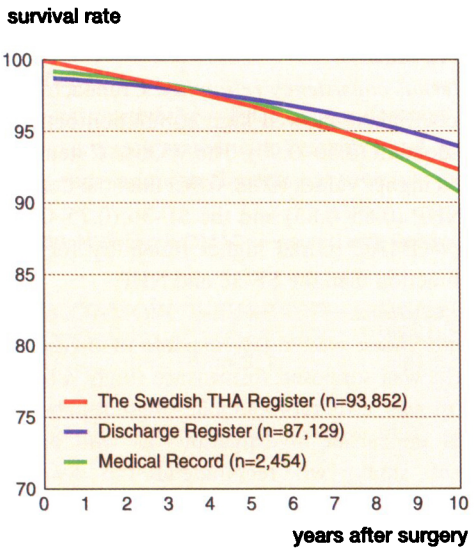

Figure 6. Survival rate for the cohort from the Swedish THA register, the Discharge register and the randomly selected cohort (medical records).

Table 4. Comparison between primary hip replacements and normal population concerning risk of death. Observed and expected number of deaths

	No. of deaths Observed	Expected	Obs/Exp	95% CI	
Men	201	88	2.28	1.97–2.61	Rheumatoid arthritis
	336	147	2.28	2.05–2.54	Hip fracture
	2,958	3,479	0.85	0.82–0.88	Arthrosis
Women	408	176	2.32	2.10–2.56	Rheumatoid arthritis
	1,054	616	1.71	1.61–1.82	Hip fracture
	2,255	2,975	0.76	0.73–0.79	Arthrosis
All diagnoses	4,036	4,005	1.01	0.98–1.04	Men
	4,889	4,597	1.06	1.03–1.09	Women

Table 5. Comparison between hip revisions and normal population concerning risk of death. Observed and expected number of deaths

	No. of deaths Observed	Expected	Obs/Exp	95% CI	
Men	35	12	3.04	2.12–4.23	Rheumatoid arthritis
	72	36	2.00	1.57–2.52	Hip fracture
	166	182	0.91	0.76–1.06	Arthrosis
Women	32	12	2.60	1.78–3.67	Rheumatoid arthritis
	109	71	1.53	1.26–1.84	Hip fracture
	99	106	0.93	0.76–1.13	Arthrosis
All diagnoses	715	669	1.07	0.99–1.15	Men
	616	565	1.09	1.01–1.18	Women

sponding figures for revision were 7 and 9% higher respectively (Table 5). The risk of death increased with follow-up time compared to a normal population for patients with hip fractures and rheumatoid arthritis subjected to total hip replacement (Figure 7).

Patients with osteoarthrosis subjected to with primary THR had lower mortality within all cause-of-death groups, except for diseases in the musculoskeletal system according to the ICD system. For hip fractures and rheumatoid arthritis, the risk increased in all diagnosis groups (Table 4). Patients with an index diagnosis of osteoarthrosis subjected to with revision hip replacement also had a significantly lower mortality, except for diseases in the musculoskeletal system, infections (men), blood and organs associated with the blood (men), tumours (women) and diseases in urinary organs (women). For revision procedures with an index diagnosis of hip fracture and rheumatoid arthritis, the risk increased, except in the tumour

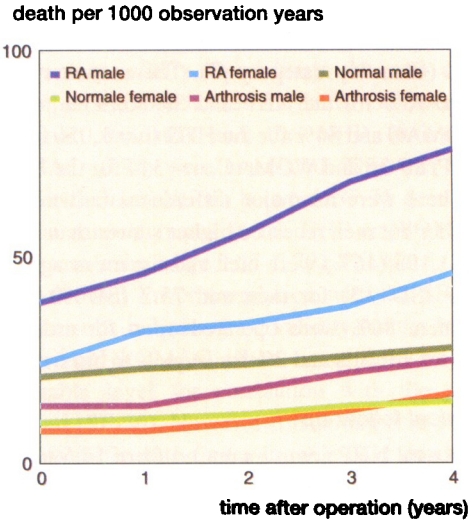

Figure 7. Incidence of death per 1000 observations year. Poisson model of a 65-year-old patient subjected to total hip arthroplasty compared to a normal person.

group, where rheumatoid arthritis patients had a lower mortality (Table 5).

Conclusion. Based on this study, the Swedish National Total Hip Arthroplasty Register seems to be valid and the annual reports are reliable. However, regular validation tests of the register are important to ensure that the high quality that the register has today is maintained.

Papers III–V: Outcome after total hip arthroplasty, part I, II and comparison of different measurement methods

The aims were, firstly to analyze the clinical outcome, general and disease-specific, in patients with primary total hip replacement and thereby test the adequacy of the end-point for failure in the Swedish THA register, and secondly to describe how total hip replacement can be followed-up in routine clinical practice, as well as, in specific research projects by comparing different outcome measurement methods.

General results

The mean age at surgery and follow-up was 69 (29–95) years and 75 (33–97) years, respectively, n=2,604. 31% of the total group had disability in one hip (Charnley category A), 18% in both hips (Charnley category B). The remaining 51% had a general disease or another disease that impaired gait (Charnley category C). The response rates were 96% for the NHP and SF-36, 93% for the WOMAC and 84% for the HHS (n = 1,056 for the NHP, SF-36 and WOMAC, n = 344 for the HHS).

There were no major differences between the gender but men received higher scores than women. 1,198 (46%) were men and the mean age was 75.7 (SD 9.3) for men and 75.2 (SD 10.6) for women. 86% were operated upon for arthrosis, 3% for arthritis and 2% for sequels to hip fracture.

Clinical follow-up

The total NHP score increased from 14.6 to 24.4 up to 10 years postoperatively and SF-36 the total score decreased from 69.7 to 59.7 during the same period. The standard deviation for domain and total score (95% confidence interval) was 7.3–38.6 for the NHP and 20.9–43.9 for the SF-36. The median total Harris Hip Score for patients operated upon with the Charnley prosthesis was 96 (range 39–100, n=32), compared to 92 for those given the Lubinus SPII prosthesis (range 34–100, n=73), and 87 for those given the Scan Hip (range 35–100, n=57). As expected, the mean total score for the Harris Hip Score and the WOMAC declined with increased follow-up time. The domain and total score in the Harris Hip Score were higher than the WOMAC scores. The clinical investigations (Harris Hip Score) and the postal survey (NHP, SF-36 and WOMAC) showed large differences between patients with one affected hip (Charnley A, total median Harris Hip Score of 96 points, range 37–100) and patients with general diseases that disabled the patients (Charnley C, total median Harris Hip Score of 79 points, range 34–98). To make it easier to compare the results for the NHP, SF-36, WOMAC and HHS, the scores were transformed to a 0–100-point scale. 100 points indicate best health (Figure 8 and 9).

The Swedish THA register has shown a 93% survival rate, on a national level, after ten years during the same period as this study. The survival based on an arbitrary estimated clinical failure was dependent on the level for clinical failure for each scoring system where the general health score (SF-36 and NHP) showed a 35 to 57% 10-year survival. Patients that were revised or scored lower than 60 points in the Harris Hip Score had an 87% 10-year survival rate in this study. For the WOMAC, the corresponding result was 80%.

There was no difference in general or disease-specific health and 10-year survival between regional, county and rural hospitals according to the NHP and SF-36 analysis. Blood loss (mean 800 mL) and operation time (mean 1 h 46 min) did not affect the 10-year result.

Radiographic follow-up

One hospital did not want to provide material for the radiographic examinations, and patients who did not completely reply to the Harris Hip Score and WOMAC were not included in the statistics for the radiographic results. 76% of the patients were examined using radiographic analysis. The total number of operations for each hospital type was low; regional and rural hospitals each had a radiographic failure rate of 9% (5 failures out of

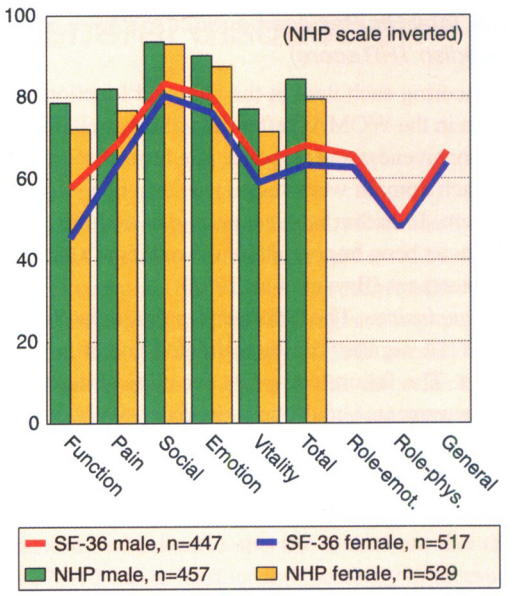

Figure 8. SF-36 and NHP results for male and female patients at follow-up after 2–10 years.

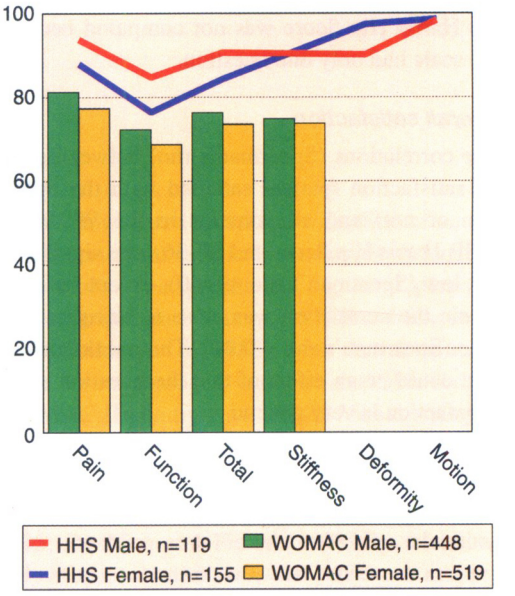

Figure 9. HHS and WOMAC results for male and female patients at follow-up after 2–10 years.

53 primary operations in regional hospitals, and 7 failures out of 79 primary operations in rural hospitals). The county hospitals had a 23% failure rate (15 failures out of 65 primary operations).

Radiographic failure was studied for the cup and/or the stem. With the exception of pain measured by the HHS, the results did not show any significant difference (on the 5% level) in pain, function or total score in the two clinical investigations between the failure and the non-failure groups.

Comparison of different measurement methods

Content validity. The content of the WOMAC and the Harris Hip Score is closely related (pain, function and stiffness), while the SF-36 and NHP contain additional general domains. Very few floor values were seen, with no floor effects at all for the disease-specific questionnaire, but there were several domains that contained ceiling values. The pain domain in the Harris Hip Score showed a very high median value and ceilings.

Construct validity. The disease-specific questionnaires (WOMAC and Harris Hip Score) gave smaller differences in total score between Charnley categories, gender, different follow-up times and age groups as than the general questionnaires. The function domain in the Harris Hip Score correlated better with function in the WOMAC, NHP and SF-36 than with pain in the WOMAC, NHP and SF-36. The same results were obtained when comparing the domains of pain in the three scores with function domains, with the exception of the SF-36. That is to say the NHP, SF-36, WOMAC and HHS showed high convergent and divergent validity.

Criterion validity. Spearman's rho was over 0.40 when the total scores for the SF-36 (used as the gold standard) and the NHP (0.66) or WOMAC (0.71) were correlated. This was also true for the correlations between SF-36 domains and the same domains in the NHP or WOMAC (Spearman's rho range 0.67–0.79). The Harris Hip Score showed lower correlation with the SF-36. These results indicate good validity for the NHP, SF-36 and WOMAC when using 0.40 as the acceptable level for correlation with the gold standard (SF-36). However it is an important fact that it is difficult to find a suitable gold standard, especially for disease specific-instruments.

Internal consistency reliability. Cronbach's alpha coefficient showed high reliability for pain and function (0.80–0.97). The WOMAC test received the highest values. The pain domain in

the Harris Hip Score was not computed because the scale had only one question.

Overall satisfaction

The correlations (Spearman's rho) between overall satisfaction (patient satisfied with the operation or not) and all domains in the WOMAC, NHP, Harris Hip Score and SF-36 were significant but low (Spearman's rho < 0.40), except for pain, where the correlation was close to being acceptable (Spearman's rho = 0.40). This weak correlation could be an effect of that the question about satisfaction is very general.

The Total Hip Replacement Score (the Swedish THR score)

By scoring each item in the pain and function domain in the WOMAC after the validity, reliability, responsiveness and response rate tests, three items in each domain were suggested for a new scoring system. In earlier studies, pain, gait and hip flexion have been important for the outcome after hip replacement (Bryant et al. 1993).

Conclusions. The failure end-point in the Swedish THA register can be validated and it is very exact. The failure end-point for clinical outcome measurements should be more extensively evaluated to provide reference material for different studies on THR, such as validations of the register by survival analyses.

General discussion

This study showed that the results from the Swedish THA register were reliable and revision was a useful end-point for failure. The information provided from the Swedish THA register is important to continue the high quality of THR surgery on a national level and thereby fulfill the criteria for evidence-based medicine. The clinical outcome of THR, from the patient's viewpoint and from the doctor's viewpoint was good, with high survival of the implant and the patients. The study also showed that a disease-specific instrument like the WOMAC has at least as high accuracy in measuring the results of hip replacement (high validity) and as good reliability (high reliability) as a general scoring system like the SF-36 and NHP. These scores can be used to study patients subjected to THR.

Previous works by the authors and by others have indicated high validity, reliability and responsiveness for the studied health evaluation instruments (SF-36, NHP, WOMAC). Disease-specific measures focus on the disorder under consideration and the patient's problems related to it. Disease-specific instrument may therefore be more relevant to the patient and the physician than general instruments, as well as better in detecting the effect of the treatment. General measures detect complications or side effects in areas of function or organ systems not specifically related to the disease under consideration. An advantage with general instruments also is the possibility to compare various medical treatments (Bombardier et al. 1995, Keller et al. 1993). The American Academy of Orthopaedic Surgeons and the Société Internationale de Chirurgie Orthopèdic et de Traumatologie recommend that a disease-specific measure, such as the WOMAC, be included in all studies of the outcome of hip arthroplasty (Laupacis et al. 1993). Disease-specific scales are more responsive than general health status measures for evaluation of the outcomes of orthopaedic procedures and the disease-specific scales correlates also to objective data (Boardman et al. 2000). However, a complete evaluation of an operative treatment such as hip arthroplasty requires the use of both a specific and a general questionnaire (Wright and Young 1997, Rorabeck et al. 1994, Bellamy et al. 1988). Both the Likert score and a visual analogue scale are available for the WOMAC (Bellamy et al. 1992). Visual analogue scales are reliable but difficult to score and cannot be administered by telephone. They are not understood by 7–10 % of the population (Katz et al. 1995, De Nies and Fidler 1997). By telephone follow-up evaluation, one can retrieve a missing answer in a single Likert question and therefore get a higher response rate (McGrory et al. 1997). With an older population, some individuals may have difficulties answering the items in the SF-36, which results in a low response rate (Dorman et al. 1997, Jenkinson 1991).

For routine clinical follow-up, several authors claim that very short general health scoring system like the EuroQol could replace the SF-36 (Brazier et al. 1993, Brazier et al. 1996, Hollingworth et al. 1995). The problem with the EuroQol is low responsiveness, but an advantage is a high response rate compared to the SF-36. How short can a questionnaire be then? Can a single item replace a domain of several items concerning the same area of interest? From a statistical point of view, the answer is no, because the single item cannot be controlled and if the patient misunderstands the question the result from the questionnaire can be misleading (Chronbach 1951, Bellamy 1993). The SF-36 is a more reliable tool for studies that require high sensitivity, which is often the case in research evaluations where the material is small. The EuroQol could be used for routine clinical practice when longitudinal long-term documentation of THR outcome is needed (Anderson et al. 1993, Brazier et al. 1993, Brazier et al. 1996, Hollingworth et al. 1995, Richards and Irwing 1997). The disease-specific Harris Hip Score has been criticised for being used only as a total score and for its low reliability in items in the

domain of motion (Arndt et al. 1998, Jacobsson et al. 1990, Liberman et al. 1996, McGrory et al. 1996, Rothwell et al. 1997, Söderman et al. 2000).

The Swedish National Total Hip Arthroplasty Register has been validated through feedback between the clinics and the register (Ahnfelt 1986, Herberts et al. 2000). Validation is also performed by means of retrospective controls through hospital medical files regarding number of procedures and also by comparison with other national databases (Havelin et al. 1993, Herberts and Malchau 1997). These validations of the register are not specific and it is difficult to judge how valid the results are based on these observations. In a survival analysis of total hip replacement, the result from 410 prospectively studied patients were compared to the Swedish THA register (Garellick et al. 2000). The study showed that a prospective procedure could result in an increased revision rate compared to observational studies like a register study because of lack of clinical and radiographic results. The ten-year survival of the implant was 89% but the clinical failure was even worse. One problem with this comparison with the register was demographic differences (such as diagnosis and age) between the 410 patients in the study and the national cohort in the register.

Another national register, the Swedish Knee Arthroplasty register, started in 1975 (Bauer et al. 1980; http://www.ort.lu.se@knee/). 80% of the knee arthroplasties performed in Sweden have been included in the register and validation studies of the register has revealed that 94% of the knee revisions were correctly registered (Knutson et al. 1994, Robertsson et al. 1999). The material in the validation study from the knee register was from the register primary and revision cases. By comparison with the Swedish Discharge register the missing patients were detected (Robertsson et al. 2000).

Concerning radiographic follow-up, some authors report no or low correlation between clinical and early radiographic failure (Boeree and Bannister 1993, Fender et al. 1999, Gustilo and Pasternak 1988, Kwong et al. 1992, Maloney et al 1990, Wixson et al 1991). Longitudinal radiographic examinations are, however, important for periprosthetic bone loss that could lead to substantial bone loss with a high risk of fracture and difficulties in revision (Garellick 1998, Malchau 1995).

One of the most important issues when planning this study was the patient selection. The patients were selected by a randomised cohort from the Discharge register. This random selected cohort were primary related to the Discharge register and the main aim was to make comparison with the Swedish THA register. The patients were random selected from the whole country and for the clinical and radiographic evaluations the patients were stratified with respect to age and gender in three geographic regions in Sweden. This selection provided patients who were not operated upon in the authors' hospital (no surgical bias) for the study and the number of patients required was calculated by power-analyses by professional medical statisticians. Hence, the material was representative for the whole nation, which made comparison with the Swedish THA register possible.

The practical parts of the study were performed within a short time so that patiens' status did not change, making it possible to compare the results between each project. A specially designed database was used to see which patients did not reply, making it possible to send reminder letters very fast. Despite this, one problem was that a few patients died during the study. There was a high follow-up rate for the different projects except the radiographic follow-up however. Although the response rate for the radiographic follow-up was acceptable, the material was too small to provide information on how to predict failure using radiographic analysis compared to the register and clinical results. Also, the patients were examined radiographically only once, which made it difficult to analyse failure.

All instruments used in clinical follow-up should be evaluated in different ways to insure high validity, reliability and responsiveness (Streiner and Norman 1989). The choice of scoring systems in this study were based on whether the systems were used worldwide and reported and validated in several countries. A new disease-specific and general self-administered scoring system were suggested in this study but even this instrument should be extensively validated before routine use. It is highly desirable to have reference material stratified for age, gender, Charnley categories for all

score systems. An important problem using these instruments is the lack of international consensus on which scoring system should be used.

By means of logistic regression, survival analyses were performed in this study to compare the clinical results with the Swedish THA register. One problem in these calculations was to decide the level for clinical failure since there are few or no guidelines in the literature. The results of the THR are dependent on implant design and surgical technique and case mix including gender, diagnosis, comorbidity and Charnley category. These findings were confirmed in the present study and the results were highly related to which evaluation systems that were used. As an example, patients with multiple joints affected, such as those with Rheumatoid arthritis, had lower scores in the SF-36, NHP, WOMAC and HHS than patients with disability in one hip. This means that a more extensive study should be performed if one wishes to evaluate the register with survival analysis based on logistic regression and using clinical score systems. On the other hand, the main reasons for hip replacements in the register are arthrosis and there are relative few patients with Rheumatoid arthritis making these types of evaluations acceptable.

The results also showed that the rate of combined clinical and radiographic failure was at least twice as high as the register presents in the survival analyses. Hence, with the knowledge that there is a difference between the revision rate presented by the register and clinical results, the strict definition of failure in the register is useful as an endpoint for primary hip replacement surgery. The hypothesis was thus verified.

People not registered as patients with hip disabilities (normal populations) have been evaluated with general (SF-36) and disease-specific (HHS) instruments (Sullivan and Karlsson 1994, Brinker et al. 1996). An age and gender matched cohort (compared to this study) from these normal populations received a total score for domains in the SF-36 of 65–90 points and for total HHS 90-95 points. The clinical outcome in this study confirmed that patients subjected to THR had nearly as good health as these normal populations. The clinical results from the different groups mentioned in the study showed that a time-dependent failure of primary hip replacement was obvious. Therefore, the practical implication of the study is that regular follow-ups are appropriate for young and active patients and for high-risk patients such as those operated upon with a new technique and most undergoing revision surgery. Appropriate follow-up intervals might be directly postoperatively, one year after the operation and then every five years for an indefinite time. However, older and less active patients operated upon with well-documented methods need less frequent follow-ups and a one-year radiographic examination without indication of failure as well as a good clinical outcome will, in most cases, be sufficient.

Conclusions

This study showed that the Swedish THA register captured 94% of the revisions performed in Sweden. Compared to the results in this study and the results based on the Discharge register, the Swedish THA register is reliable.

1 The clinical and radiographic failure rates (revision and impending failures) are in several tests at least twice as high as in the Swedish THA register (revision only). The clinical results are, however, dependent on demographics, the definition of clinical failure and the scoring system used, making the results presented by the register very exact but with limited clinical value.

2 The clinical result of THR was very good both from a self- and staff-administered point of view in all hospital types, with low mortality. These results based on a national level should encourage continuous high quality THR surgery by the means of evidence-based medicine.

3 The most optimal way to follow-up the results of THR is of course that the surgeon sees his/her patients to answer any questions. The patients subjected to THR can also be followed up with a self-administered scoring system such as a disease-specific instrument (WOMAC) and a general instrument (SF-36). However, for some patients (routine clinical follow-up) a shorter questionnaire such as the Swedish THR score is desirable. A short questionnaire would also be a good complement to the register, providing reference material for further analysis of the register. The choice of clinical scoring system must be agreed by international consensus. Clinical follow-up should be supplemented by radiographic follow-up since there is a low correlation between impending radiographic failure and clinical failure.

Acknowledgments

I wish to express my sincere gratitude to:

Associate Professor Henrik Malchau and Professor Peter Herberts, my tutor and co-tutor, for guidance in the scientific field and into the field of hip replacement surgery.

Professor Björn Rydevik, Chairman, Department of Orthopaedics, Göteborg University and Professor Tommy Hansson, Head of Orthopaedics, Sahlgrenska University Hospital, for providing scientific conditions and making high quality clinical work possible.

My co-authors Olof Johnell, Hans Regnér, Göran Garellick and Roland Zügner.

My colleagues and friends at the knee register in Lund for support and in Göteborg for friendship, support and practical jokes (?). Thank you, Johan Uvehammer for the statistical manual.

Anders Odén for statistical advice. Roger Salomonsson and Albert Andinsson for programming and database design.

Marie Hagman, Karin Lindborg, Catarina Sporre and Anna Kajsa Erikson for all paper exercise (when will it end?).

All involved patients and staffs in Sweden for support.

Professor Anders Rydholm and Associate Professor Kaj Knutson for careful editorial work during the preparation of this thesis as a supplement to Acta Orthopaedica Scandinavica.

The Volvo Research Foundation, the Swedish National Board of Health and Welfare, the Gothenburg Medical Association, the Dr Félix Neubergh Foundation, the Greta & Einar Asker Foundation and the Hjalmar Svensson Foundation for financial support.

But above all—my wife Carina who has "put up with me" for 20 years. But next year....

References

Amadio P C. Outcome measurements. J Bone Joint Surg (Am) 1993; 75 (11): 1583-4.

Ahnfelt L, Andersson G, Herberts P. Re-operation av totala höftledsplastiker. Läkartidningen 1980; 77 (30-31): 2604-7.

Ahnfelt L. Re-opererade totala höftledsplastiker i Sverige under åren 1979-1983. Thesis. Göteborg University, Göteborg, Sweden, 1986.

Ahnfelt L, Herberts P, Malchau H, Andersson G B J. Prognosis of total hip replacement. A Swedish multicenter study of 4,664 revisions. Acta Orthop Scand (Suppl. 238) 1990; 61: 1-26.

Anderson R T, Aaronson N K, Wilkin D. Critical review of the International assessments of health-related quality of life. Qual Life Res 1993; 2: 369-95.

Andersson G. Hip assessment: a comparison of nine different methods. J Bone Joint Surg (Br) 1972; 54 (4): 621-5.

Arndt D C, Mohamed N N, McGrory B J, Harris W H. Validation of the Harris Hip Score as a self-administered questionnaire. Poster at the 44th annual meeting, Transactions ORS, New Orleans, Louisiana 1998; 23: 420.

Bauer G, Knutson K, Lindstrand A. Swedish knee arthroplasties. Läkartidningen 1980; 77 (22): 2088-91.

Bellamy N, Buchanan 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 and the knee. J Rheumatol 1988; 15: 1833-40.

Bellamy N, Campbell J, Stevens J, Pilch L, Stewart C, Mahmood Z. Validation study of a computerized version of the Western Ontario and McMaster Universities VA3.0 Osteoarthritis Index. J Rheumatol 1997; 24 (12): 2413-5.

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.

Bellamy N, Wells G, Campbell J. Relationship between severity and clinical importance of symptoms in osteoarthritis. Clin Rheumatol 1991; 10 (2): 138-43.

Bellamy N. Musculoskeletal clinical metrology. Kluwer Academic Publishers, Dordrecht 1993.

Black N. Why we need observational studies to evaluate the effectiveness of health care. BMJ 1996; 312: 1215-18.

Boeree N R, Bannister G C. Cemented total hip arthroplasty in patients younger than 50 years of age. Clin Orthop 1993; 287: 153-9.

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

Boardman D L, Dorey F, Thomas B J, Lieberman J R. The accuracy of assessing total hip arthroplasty outcomes. A prospective correlation study of walking ability and 2 validated measurement devices. J Arthroplasty 2000; 15 (2): 200-4.

Brazier J E, Harper R, Jones N M B, 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: 160-4.

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

Brazier J E, Walters S J, Nicholl J P, Kohler B. Using the SF-36 and EuroQol on an elderly population. Qual Life Res 1996; 5: 195-204.

Breslow N E, Day N E. The Poisson assumption. In: Statistical methods in cancer research. Vol. 2: The design and analysis of cohort studies. IARC Sci Pub, Lyon 1987: 131-5.

Brinker M R, Lund P J, Cox D D, Barrack R L. Demographic biases found in scoring instruments of total hip arthroplasty. J Arthroplasty 1996; 11 (7): 820-30.

Bryant M J, Kernohan W G, Nixon J R, Mollan R A B. A statistical analysis of hip scores. J Bone Joint Surg (Br) 1993; 75 (5): 705-9.

Bulstrode C. Total hip replacement: the way forward. Ann R Coll Surg Engl 1996; 78: 129-32.

Carmines E G, Zeller R A. Reliability and validity assessment. Sage Publication Inc, Beverly Hills 1979.

Charnley J: Numerical grading of clinical results. In: Low friction arthroplasty of the hip: theory and practice. Springer-Verlag, Berlin, Heidelberg, New York 1979: 23-4.

Chetter I C, Spark J I, Dolan P, Scott D J A, Kester R C. Quality of life analysis in patients with lower limb ischaemia suggestions for European standardisation. Eur J Vasc Endovasc Surg 1997; 13: 597-604.

Chronbach L J. Coefficient alpha and the internal structure of tests. Psychometrika 1951; 16: 297-334.

De Nies F, Fidler M W. Visual analog scale for the assessment of total hip arthroplasty. J Arthroplasty 1997; 12 (4): 416-9.

Dorman P J, Slattery J, Farrell B, Dennis M S, Sandercock P A. A randomised comparison of the EuroQol and Short Form-36 after stroke. BMJ 1997; 315: 461.

Essink-Bot M-L, Krabbe P F M, Bonsel G J, Aaronson N K. An empirical comparison of four generic health status measures. Med Care 1997; 35 (5): 522-37.

Fender D, Harper W M, Gregg P J. Outcome of Charnley total hip replacement across a single health region in England. The results of five years from a regional hip register. J Bone Joint Surg (Br) 1999; 81(4): 577-81.

Garellick G, Malchau H, Herberts P, Hansson E, Axelsson H, Hansson T. Life expectancy and cost utility after total hip replacement. Clin Orthop 1998; 346: 141-51.

Garellick G, Malchau H, Regnér H, Herberts P. The Charnley versus the Spectron Hip Prosthesis. Radiographic evaluation of a randomized, prospective study of 2 different hip implants. J Arthroplasty 1999; 14 (4): 414-25.

Garellick G, Malchau H, Herberts P. Survival of total hip replacements. A comparison of a randomized trial and a registry. Clin Orthop 2000; 375: 157-67.

Garellick G. On outcome assessment of total hip replacement. Thesis. Göteborg University, Göteborg, Sweden, 1998.

Gross M. Innovations in surgery. A proposal for phased clinical trials. J Bone Joint Surg (Br) 1993; 75 (3): 351-4.

Gustilo R B, Pasternak H S. Revision total hip arthroplasty with titanium ingrowth prosthesis and bone grafting for failed cemented femoral component loosening. Clin Orthop 1988; 235: 111-9.

Guyatt G H, Veldhuyzen Van Zanten S J O, Feeny D H, Patrick D L. Measuring quality of life in clinical trials: a taxonomy and review. CMAJ 1989; 140: 1441-8.

Harris W H. Traumatic arthritis of the hip after dislocation and acetabular fractures: treatment by mold arthroplasty. J Bone Joint Surg (Am) 1969; 51 (4): 737-55.

Havelin L I, Espehaug B, Vollset S E, Engesaeter L B, Langeland N. The Norwegian arthroplasty register. A survey of 17,444 hip replacements 1987-1990. Acta Orthop Scand 1993; 64 (3): 245-51.

Herberts P, Ahnfelt L, Malchau H, Strömberg C, Andersson B J. Multicenter clinical trials and their value in assessing total joint arthroplasty. Clin Orthop 1989; 249: 48-55.

Herberts P, Malchau H. How outcome studies have changes total hip arthroplasty practices in Sweden. Presidential guest speaker. Clin Orthop 1997; 344: 44-60.

Herberts P, Malchau M. Long-term registration has improved the quality of hip replacement. A review of the Swedish THR Register comparing 160,000 cases. Acta Orthop Scand 2000; 71(2): 111-21.

Hodgkinson J P, Shelly P, Wroblewski B M. The correlation between the roentgenographic appearance and operative findings at the bone-cement junction of the socket in Charnley low friction arthroplasties. Clin Orthop 1988; 228: 105-9.

Hollingworth W, Mackenzie R, Todd C J, Dixon A K. Measuring changes in quality of life following magnetic resonance imaging of the knee: SF-36, EuroQol or Rosser index? Qual Life Res 1995; 4: 325-34.

Hozack W J, Rothman R H, Albert T J, Balderston R A, Eng K. Relationship of total hip arthroplasty outcomes to other orthopaedic procedures. Clin Orthop 1997; 344: 88-93.

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: 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 1981; 15A: 221-9.

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: 551-9.

Jacobsson S-A, Djerf K, Wahlström O. A comparative study between McKee-Farrar and Charnley arthroplasty with long-term follow-up periods. J Arthroplasty 1990; 5: 9-14.

Jenkinson C. Why are we weighting? A critical examination of the use of item weights in a health status measure. Soc Sci Med 1991; 32: 1413-6.

Kaplan E L, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc 1958; 53: 457-81.

Katz J N, Phillips C R, Poss R, Harrast J J, Fossel A H, Liang M H, Sledge C B. Correspondence. J Bone Joint Surg (Am) 1996; 78 (9): 1445-8.

Katz J N, Phillips C B, Poss R, Harrast J J, Fossel A H, Liang M H, Sledge C B. The validity and reliability of a total hip arthroplasty outcome evaluation questionnaire. J Bone Joint Surg (Am) 1995; 77 (10): 1528-53.

Keller R B, Rudicel S A, Liang M H. Outcome research in orthopaedics. J Bone Joint Surg (Am) 1993; 75 (10): 1562-74.

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.

Krabbe P F, Essink-Bot M-L, Bonsel G J. On the equivalence of collectively and individually collected responses: standard-gamble and time-tradeoff judgments of health states. Med Decis Making 1996; 16 (2): 120-32.

Kwong L M, Jasty M, Mulroy R D, Maloney W J, Bragdon C, Harris W H. The histology of the radiolucent line. J Bone Joint Surg (Br) 1992; 74 (1): 67-73.

Kärrholm J, Frech W, Nivbrant B, Malchau H, Snorrason F, Herberts P. Fixation and metal release from the Tifit femoral stem prosthesis. 5-year follow-up of 64 cases. Acta Orthop Scand 1998; 69 (4): 369-78.

Landy F J. Stamp collecting versus science. Am Psychol 1986; 41: 1183-92.

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.

Lieberman J R, Dorey F, Shekelle P, Schumacher L, Thomas B J, Kilgus D J, Finerman G A. Differences between patients' and physicians' evaluations of outcome after total hip arthroplasty. J Bone Joint Surg (Am) 1996; 78 (6): 835-9.

Lidgren L. The bone and joint decade 2000-2010. An update. Acta Orthop Scand 2000; 71(1): 3-6.

Malchau H, Herberts P, Söderman P, Odén A. Prognosis of total hip replacement. Update and validation of results from the Swedish national hip arthroplasty registry 1979-1998. Scientific exhibition presented at the 67th annual meeting of the AAOS, March 2000, Orlando, USA.

Malchau H, Herberts P, Ahnfelt L. Prognosis of total hip replacement in Sweden. Follow-up of 92,675 operations performed 1978-1990. Acta Orthop Scand 1993; 64 (5): 497-506.

Malchau H. On the importance of stepwise introduction of new hip implant technology. Assessment of total hip replacement using clinical evaluation, radiostereometry, digitised radiography and National Hip Registry. Thesis. Göteborg University, Göteborg, Sweden, 1995.

Maloney W J, Jasty M, Rosenberg A, Harris W H. Bone lysis in well-fixed cemented femoral components. J Bone Joint Surg (Br) 1990; 72 (6): 966-70.

Mancuso C A, Salvati E A, Johanson N A, Peterson M G E, Charlson M E. Patients´ expectations and satisfaction with total hip arthroplasty. J Arthroplasty 1997; 12 (4): 387-96.

Martin D P, Engelberg R, Agel J, Swiontkowski F. Comparison of the musculosketal 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.

McGrory B J, Harris W H. Can the Western Ontario and McMaster Universities (WOMAC) Osteoarthritis Index be used to evaluate different hip joints in the same patient? J Arthroplasty 1996; 11 (7): 841-4.

McGrory B J, Morrey B F, Rand J A, Ilstrup D M. Correlation of patient questionnaire responses and physician history in grading clinical outcome following hip and knee arthroplasty. A prospective study of 201 joint arthroplasties. J Arthroplasty 1996; 11 (1): 47-57.

McGrory B J, Shinar A A, Freiberg A A, Harris W H. Enhancement of the value of hip questionnaires by telephone follow-up evaluation. J Arthroplasty 1997; 12 (3): 340-3.

McHorney C A, Ware J E Jr, Raczek A E. 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.

McKenna S P, Hunt S M, McEwen J. Weighting the seriousness of perceived health problems using Thurstone´s method of paired comparisons. J Epidemiol 1981; 10 (1): 93-7.

Morris R W. Evidence-based choice of hip prostheses. J Bone Joint Surg (Br) 1996; 78 (5): 691-3.

Mulroy W F, Harris W H. Acetabular and femoral fixation 15 years after cemented total hip surgery. Clin Orthop 1997; 337: 118-28.

Nivbrant B. The femoral component in total hip arthroplasty. An evaluation of stem design, fixation principles and loosening scenario. Thesis. Umeå University, Umeå, Sweden, 1999.

Richards D M, Irving M H. Assessing the quality of life of patients with intestinal failure on home parenteral nutrition. Gut 1997; 40: 218-22.

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

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

Ritter M A, Albohm M J. Overview: Maintaining outcomes for total hip arthroplasty. The past, present, and future. Clin Orthop 1997; 344: 81-7.

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 1999; 70 (5): 467-72.

Robertsson O, Lewold S, Knutson K, Lidgren L. The Swedish knee arthroplasty project. Acta Orthop Scand 2000; 71 (1): 7-18.

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; 28: 210-5.

Rorabeck C H, Bourne R B, Laupacis A, Feeny D, Wong C, Tugwell P, Leslit K, Bullas R. A double-blind study of 250 cases comparing cemented with cementless total hip arthroplasty. Clin Orthop 1994; 298: 156-64.

Rothwell P M, McDowell Z, Wong C K, Dorman P J. Doctors and patients don't agree: cross sectional study of patients' and doctors' perceptions and assessments of disability in multiple sclerosis. BMJ 1997; 313: 1580-3.

Rowley D I. Outcome studies - Why bother? Orthopaedic Product News 1997; Dec 96/Jan/Feb: 34-5.

Smith-Petersen M N. Arthroplasty of the hip: a new method. J Bone Joint Surg 1939; 21(2): 269-88.

Sochart D H, Long A J, Porter M L. Joint responsibility: the need for a national arthroplasty register. BMJ 1996; 313: 66-7.

Streiner D L, Norman G R. Health measurement scales. A practical guide to their development and use. Oxford University Press, Oxford, New York, Tokyo, 1989.

Sullivan M, Karlsson J, Ware J R. 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.

Sullivan M, Karlsson J. SF-36 hälsoenkät. Svensk manual och tolkningsguide. (Swedish manual and interpretation guide.) Göteborg University and Sahlgrenska University Hospital, Göteborg, Sweden, 1994.

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.

Söderman P, Malchau H. Is the Harris Hip Score system useful to study the outcome of total hip replacement? Accepted for publication in Clin Orthop, Feb. 2000.

Söderman P, Malchau M. Outcome measurements in total hip replacement surgery (THR). In: Outcome measuring. Spris förlag, Stockholm 1998; 89-95.

Thanner J, Kärrholm J, Herberts P, Malchau H. Porous cups with and without hydroxyapatite-tricalcium phosphate coating. 23 matched pairs evaluated with radiostereometry. J Arthroplasty 1999; 14 (3): 266-71.

Uyl-de Groot C, Rutten F F H, Bonsel G J. Measurement and valuation of quality of life in economic appraisal of cancer treatment. Eur J Cancer 1994; 30A (1): 111-7.

Van Agt H M E, Essink-Bot M-L, Krabbe P F M, Bonsel G J. Test-retest reliability of health state valuations collected with the EuroQol questionnaire. Soc Sci Med 1994; 39 (11): 1537-44.

Ware J E, Sherbourne C D. The MOS 36-Item Short-Form Health Survey (SF-36). Med Care 1992; 30 (6): 473-83.

Wiklund I, Romanus B. Nottingham health profile. Livskvalitetsbedömning hjälp vid operationsprioritering. Läkartidningen 1988; 85 (38): 3060-1.

Wixson R L, Stulberg S D, Mehlhoff M. Total hip replacement with cemented, uncemented and hybrid prostheses. A comparison of clinical and radiographic results at two to four years. J Bone Joint Surg (Am) 1991; 73 (2): 257-69.

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

Öberg U. Functional Assessment System of lower-extremity dysfunction. Thesis. Linköping University, Linköping, Sweden, 1996.