Patient Reported Outcomes in Hip Arthroplasty Registries

Aksel Paulsen

This review has been accepted as a thesis together with 4 previously published papers by University of Southern Denmark 13th of December 2013 and defended on 7th of February 2014.

Tutors: Søren Overgaard, Ewa M. Roos and Alma Becic Pedersen.

Official opponents: Nils Hailer, Peter Vedsted and Jan Hartvigsen.

Correspondence: Department of Orthopaedic Surgery and Traumatology, Odense University Hospital, Sdr. Boulevard 29, 5000 Odense C, Denmark.

E-mail: akselpaulsen@gmail.com

Dan Med J 2014;61(5):B4845

THIS THESIS IS BASED ON THE FOLLOWING 4 PAPERS:

1. Paulsen A, Pedersen AB, Overgaard S, Roos EM. Feasibility of four patient-reported outcome measures in a registry setting. A cross-sectional study of 6000 patients from the Danish Hip Arthroplasty Registry. Acta Orthopaedica 2012; 83 (4): 321–327.

2. Paulsen A, Overgaard S, Lauritsen JM. Quality of Data Entry Using Single Entry, Double Entry and Automated Forms Processing - An Example Based on a Study of Patient-Reported Outcomes. PLoS ONE 2012 7(4): e35087.

3. Paulsen A, Odgaard A, Overgaard S. Translation, cross-cultural adaptation and validation of the Danish version of the Oxford Hip Score - Assessed against generic and disease-specific questionnaires Bone and Joint Research, 2012; 1 (9): 225-233.

4. Paulsen A, Roos EM, Pedersen AB, Overgaard S. Minimal clinically important improvement (MCII) and patient acceptable symptom state (PASS) in total hip arthroplasty (THA) patients 1 year postoperatively. A prospective cohort study of 1335 patients. Acta Orthopaedica 2014; 85 (1): 39–48.

The papers will be referred in the text by their Roman numerals (I–IV)

INTRODUCTION AND BACKGROUND

Historical background of Total Hip Arthroplasty

Osteoarthritis (OA) has been common in humans since Paleolithic times (8). Amputation and joint excision arthroplasty, osteotomies and pseudarthrosis, interpositional hip arthroplasty with soft tissue hip interpositions have been used as interventions - mostly unsuccessful- in the last three centuries, before the Norwegian-born American surgeon Marius Smith-Petersen in 1938 implanted synthetic molded prosthesis with good clinical results (9).

Sir John Charnley revolutionized total hip arthroplasty (THA) in the 1950s and 60s, by using acrylic cement, introducing high-density polyethylene as a bearing material, and introducing low friction torque arthroplasty. These implants had a 77-81 % implant survival at 25-year follow-up with revision of any component as the endpoint (10;11). The improvement of THA did not stop there, and the present implant survival in the large populations of the different national hip arthroplasty registries (12-19), has earned the THA the status as 'the operation of the century' (20).

THA

THA for patients with end-stage primary OA is a successful orthopedic procedure in relation to implant survival (12;21-23). THA is indicated for patients with pain, functional disabilities and reduced quality of life (24). End-stage primary OA constitutes the largest group of patients.

In Scandinavia almost 36,000 primary THA are performed each year, approximately 20,000 in Sweden, approximately 7,000 in Norway and approximately 9,000 in Denmark (19;25;26). More than 285,000 THA are performed each year in the USA (27), and almost 90,000 in the UK (28). The incidence of THA has been increasing during the last decades due to the improvements in surgical technique and ageing of the population (giving an increase in the prevalence of arthritic disease) as well as expansions of the indications for surgery (29-31).

In Denmark the incidence of primary THA in Denmark increased from 101 to 134 per 100,000 inhabitants during the period 1995 to 2002. In 2010 the incidence peaked to 160 per 100,000, but fell to 155 per 100,000 inhabitants in 2011 (26). Even though the number of THA varies from year to year, the number is expected to continue to rise, and an additional increase by 22% in 2020 is expected, based only on expected changes in age distribution (32).

Hip Arthroplasty Registries

Since the initiation of the first Hip Arthroplasty Registry in Sweden in 1979, other Nordic national hip arthroplasty registries have emerged. Since 1980, the Finnish Arthroplasty Register has been collecting information on THAs (33). The Norwegian Arthroplasty Register started registration of THAs in September 1987 (34). The DHR was established the 1st of January 1995. From the 1st of January 1995 to 31st of December 2011, 111,907 primary THA and 17,791 revisions have been reported to DHR (26). Since the establishment of DHR, many other national Hip Arthroplasty Registries has been established (15-18;28). The initial main purpose of the Hip Arthroplasty Registries was to improve the treatment of THA patients, by detecting inferior results of implants as early as possible (34). Later the focus has also included research activity; national observational studies have some noticeable advantages compared to randomized clinical trials: a large number of patients included, the possibility to perform analyses of uncommon complications, a high statistical power, and no performance bias. The Nordic Arthroplasty Register Association (NARA) was started in 2007, resulting in a common database for Denmark, Norway, Sweden and Finland with regard to hip- and knee replacements with the main target to further improve and facilitate the Nordic research concerning implant surgery. NARA aims to perform outcome analyses (in general and for specific implants), analyze patient demographics of the participating countries, construct a standardized 'case-mix indicator' to be used in comparisons, as well as to stimulate PhD students from the different countries to use the unique Nordic data in research activity. The first NARA-projects has been completed and included over 280,000 THAs (35). In parallel with the increased number of Hip Arthroplasty Registries, the value of arthroplasty registry data has become increasingly clear (36;37).

Traditional Outcome Measures

Traditional endpoints in studies among THA patients are mortality and morbidity rates, operative complications (intraoperative fractures, superficial or deep wound infections, deep venous thrombosis, pulmonary embolism and postoperative dislocation) and the lifetime of the prosthetic materials before implant failure. Seen from a patient perspective a prosthesis still in place may not be the correct definition of surgery success; pain, physical function and quality of life is of more importance (38-41). There seems to be one or more subgroups of patients who do not benefit from the surgery due to persistent pain and/or functional limitations. In the Swedish Hip Arthroplasty Register, 14% of the patients were not satisfied after the THA (19). In Denmark 6% of primary THA patients were 'unsatisfied' or 'not completely satisfied' minimum six months postoperative according to the 2005 annual report (42). Other reports show that 10-15% of patients report persistent pain and functional limitation postoperatively (43), and 14-36% of patients do report that they have not benefitted from the operation (44), making implant survival alone a suboptimal success criterion.

Outcome has been assessed based on patient survival, implant survival, the amount of joint pain and the postoperative joint function. Joint function (by number of degrees in hip flexion, rotation, adduction and abduction) has been used to measure the success of THA and are included in the Harris Hip Score (45). But the number of degrees in hip motion alone is no precise measure of success (and only constitutes a small amount of points in Harris Hip Score), as a low number of degrees in hip motion alone only represents a minor fraction of a patient's functional disability – one of several indications for THA. The assessments have traditionally been made by the surgeon. Hip scores, like Charley's modification of the Merle d'Aubigné and Postel score (46) and the original Harris hip score, were created as a mean to summarize clinical and radiological data, to better describe the postoperative situation and current hip status. These scores were surgeon-based hip scores, where the surgeon assessed the patient's amount of pain and the patient's physical function after talking to the patient and doing a clinical examination (although 37 of 773 of Charley's patients actually self-reported due to that they were living far away) (46). Inclusion of these endpoints (presence of severe pain, low functional scores, and radiographic evidence of loosening) do not give any information on patient's satisfaction or health-related quality of life (HRQoL). Since it can be substantial disagreement between doctors and patients about health status (1;47), and it is the patients perspective of pain, HRQoL and physical function that is main importance as an indication for THA today, it is clear that patient reported outcomes (PRO)s is the best way to assess the postoperative result of THA.

Patient Reported Outcome Measures

The desire to find a better measure of success has motivated the clinicians to focus on PROs to be used in national clinical databases (48-52). In the past few decades several new PROs have been introduced and used in research. Since 2006 the US Food and Drug Administration has strongly recommended the inclusion of PROs in clinical trials (53;54) and PROs are increasingly being introduced in national hip arthroplasty registries (55-58). The Department of Health in England now requires the routine measurement of PROs for all National Health Service patients in England before and after they undergo total knee arthroplasty or THA (59), and the Swedish Hip Arthroplasty Registry introduced a PRO follow-up program as a pilot project in 2002, which has now been adopted by nearly all units performing THA in Sweden (55).

In addition to the possibility of gaining access to the patient perspective of THA without an external interpretation, PROs may also have better reliability and validity than some clinical measures. The reliability reported for the OHS items (Paper III, Table 7) was comparable to hip muscle force measurement reliability in patients older than 65 years (60). The ICC reported for the OHS items (Paper III, Table 7) was better than the reported goniometer ICC measuring hip range of motion (61). In hip fracture patients, the responsiveness of performance-based measures was higher than for PRO measures for mobility, but not for balance or strength (62). Latham et al. conclude that the validity, sensitivity, and responsiveness of PRO measures of physical function are comparable to performance-based measures after hip fracture, and that both measures would be suitable in clinical trials examining improvement in physical function (63).

With the increased focus on- and usage of PROs, it has become more important to establish quality criteria for measurements properties of PROs (for example; the construct validity is adequate if hypotheses are specified in advance and at least 75% of the results are in correspondence with these hypotheses, in (sub)groups of at least 50 patients) (64), and also to establish agreement on definitions and taxonomy of their measurement properties (3;65). Developing PROs to meet these quality criteria is very time consuming, and translating PROs from a source language to another additionally gives the possibility of international comparisons, if done correctly (66).

The increased focus on measuring and validating measurement tools (67), and on PROs, has also lead to an increased interest on how to interpret PRO results (68). In registry settings with a high number of patients included, differences in PRO scores or change scores may often be statistically significant. However, this does not express that the patient have had a relevant improvement. Thus unless minimal clinically important improvement (MCII) and patient acceptable symptom state (PASS) have been estimated, postoperative PRO scores and change scores have unknown clinical relevance, and PRO results may be very difficult to interpret.

Hip-specific PROs, and PROs concerning general health

The general overview of some of the PROs that has been used for THA patients is presented in the Table 1. The PROs can be divided into disease/site-specific and those concerning general health (generic). The included disease/site-specific PROs will be referred to as hip-specific PROs. There are good reasons to use both hip-specific PROs and PROs concerning general health -while the first are specially designed to be relevant to a narrow patient group and may shed light on specific problems THA patients have postoperative, the latter may give more information on general health issues of importance for the outcome. The PROs provides numerical endpoints, e.g. one or more sum scores, which define the clinical outcome. These PROs do not provide information about what is important to the individual patient, or if the patients preoperative expectations have been met. Work is done to develop and validate personalized scoring systems to assess the individual effect of disability in patients with OA (69), and to identify main concerns of the patients (70). PROs and items regarding patient satisfaction may be affected by factors unrelated to the surgical intervention itself, such as the patient-surgeon relationship and the process of care (71), making the patient satisfaction a problematic outcome to interpret.

Table 1. PROs used for THA patients

PROs	
Hip Specific	**McMaster Toronto Arthritis Patient Preference Disability Questionnaire (MACTAR)** **Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC)** **Hip dysfunction and Osteoarthritis Outcome Score (HOOS)** **Oxford Hip Score (OHS)** **Arthritis Impact Measurement Scales (AIMS)** **Forgotten Joint Score-12 (FJS-12)**
General Health (generic)	**Nottingham Health Profile (NHP)** **Sickness Impact Profile (SIP)** **Medical Outcomes Study 36-Item Short-Form Health Survey (SF-36)** **Medical Outcomes Study 12-Item Short-Form Health Survey (SF-12)** **EuroQol-5D-3L (EQ-5D)**

Measurement properties

The measurement properties of a PRO are of paramount importance. Validity is defined as the degree to which a PRO instrument measures the construct(s) it purports to measure. It includes content validity (including face validity), construct validity (including structural validity, hypothesis-testing, cross-cultural validity) and criterion validity (including concurrent validity, predictive validity). Content validity is defined as the degree to which the content of an HR-PRO instrument is an adequate reflection of the construct to be measured. Face validity is defined as the degree to which (the items of) an HR-PRO instrument indeed looks as though they are an adequate reflection of the construct to be measured. Construct validity is defined as the degree to which the scores of a PRO instrument are consistent with hypotheses (for instance with regard to internal relationships, relationships to scores of other instruments, or differences between relevant groups) based on the assumption that the PRO instrument validly measures the construct to be measured.

Structural validity is defined as the degree to which the scores of an HR-PRO instrument are an adequate reflection of the dimensionality of the construct to be measured. Cross-cultural validity is defined as the degree to which the performance of the items on a translated or culturally adapted HR-PRO instrument is an adequate reflection of the performance of the items of the original version of the HR-PRO instrument. Criterion validity is defined as the degree to which the scores of an HR-PRO instrument are an adequate reflection of a 'gold standard' (3). Strauss and Smith highlights five recent advances in validation theory and methodology of importance for clinical researchers, among them an increasing appreciation for theory and the need for informative tests of construct validity, in their review exploring the history of validation efforts (72). Quality criteria for content validity, construct validity and criterion validity have been proposed (64). In addition to face validity, construct validity by hypothesis testing was assessed for OHS in study III. Factor analysis was used to examine the dimensionality of all PROs or PRO subscales included in study I.

Reliability is defined as the extent to which scores for patients who have not changed are the same for repeated measurement under several conditions. It includes internal consistency (the degree of the interrelatedness among the items), reliability (including test-retest, inter rater, intra rater) and measurement error (including test-retest, inter rater, intra rater) (3). Quality criteria for internal consistency, and reliability have been proposed (64). Test-retest reliability and internal consistency were assessed for OHS in study III. Reliability will be further covered by the paragraphs on distribution based measures of change in the methodological considerations, and in the discussion.

Responsiveness is defined as the ability of an HR-PRO instrument to detect change over time in the construct to be measured (3). Quality criteria for responsiveness have been proposed (64). Two main approaches can be used for assessing responsiveness, the criterion approach (in situations where there is a gold standard for the construct to be measured) and the construct approach (in situations where there is no gold standard for the construct to be measured). In situations where an original PRO and a short version of this PRO are used, the original PRO can be considered to be a gold standard. Otherwise gold standards in PRO research are rare. A five point global rating scale (a single follow-up question concerning change since baseline) can be considered a reasonable gold standard if it assesses the same construct as the PRO (73). In study IV a five point global rating scale concerning change in hip problems was used. The construct in the PROs used in study IV were hip pain (HOOS Pain), hip function (HOOS-PS), hip related quality of life (HOOS QoL), general mobility (EQ-5D question 1), general self-care (EQ-5D question 2), general usual activities (EQ-5D question 3), general pain/discomfort (EQ-5D question 4), general anxiety/depression (EQ-5D question 5) and general current state of health (EQ-VAS). Thus the responsiveness of HOOS and EQ-5D was assessed with a construct approach.

Interpretability is defined as the degree to which one can assign qualitative meaning - that is, clinical or commonly understood connotations – to an instrument's quantitative scores or change in scores (3). Quality criteria for interpretability have been proposed (64). I have reported distributions of PRO scores in study III (Paper III, Figure 2) and in study IV (Paper IV, Supplementary data, Figure 2). Floor and ceiling effects are reported in study I (Paper I, Table 3) and study III (Paper III, Table III). MCII and PASS has been reported in study IV (Paper IV, Table 2 and Table 3). PASS for subgroups have been reported in study IV (Paper IV, Table 4).

The content validity, internal consistency, criterion validity, construct validity, reproducibility (agreement and reliability), responsiveness, interpretability and floor and ceiling effects should be documented and acceptable (64), as further outlined in the methodological considerations. To be able to effectively communicate findings to the rest of the research community, a consensus on definitions and taxonomy describing measurement properties is emerging (3;65).

Data quality

Using PRO data have several potential pitfalls for errors. Reider and Lauritsen point out some of these potential errors, arising from data capture, poor design of the data entry form, no program constraints on data entry, single-entry manual key punching and lack of validation, in the table from their work (74). Automated form processing (AFP) may streamline and improve the process and potentially improve the data quality.

Data collecting, data handling and document processing

Research on document processing began in the 1960s (75-81). With the rapid development of modern computers and the increasing need to acquire large volumes of data, automatic text segmentation and discrimination research took off in the early 1980s (82-84). The rapid evolution in software and hardware for automated forms processing, have led to a wide variety of devices and technologies available today to collect subjective data including different kinds of AFP scannable forms (85-87). In the AFP process one 'automatically' capture information from data fields by scanning, and convert these data into an electronic format. A template contains details on where the data fields are located within the form or document, like a 'map' of the document. The data are then recognized automatically using the pre-defined templates and configurations, but verification by a human operator is required if the program is uncertain.

Despite the rising in usage of PROs, and the increasing amount of data acquired in the health services, paper forms are still often used to capture PROs. Paper forms may often be the chosen way of administration, especially when dealing with an elderly population, as it is known that some patient groups does not respond adequately to an Internet-based application for collecting PROs (55). For transferring data to an electronic format, manual double entry of data is still defined as the definitive gold standard of Good Clinical Practice (88). But the manual double-key entering of data by key punching is laborious, costly and can give a grave reduction in data quality, if the proportion of erroneous entries is big (89;90). Document processing by AFP has been suggested as an alternative.

AIMS

The main objectives of the work presented in this PhD thesis were:

PAPER I

To determine the feasibility of four PROs, including the EQ-5D, the SF-12, the HOOS, and the OHS, by testing response rate, floor and ceiling effect, missing items, and need for manual validation of forms among THA patients registered in the DHR. I also aimed at calculating the number of patients needed to discriminate between subgroups of age, sex, diagnosis, and prosthesis type for the EQ-5D, the SF-12, the HOOS, and the OHS in a hypothetical repeat study.

PAPER II

To examine and validate an up-to-date AFP system, by comparing paper-based and scanned PRO forms with single and double manually entered data.

PAPER III

To develop an adequately translated and culturally adapted Danish language version of the OHS for use in the DHR.

PAPER IV

To find cut-points for the minimal clinically important improvement based on changes in PRO scores and the acceptable postoperative PRO score, by estimating MCII and PASS 1 year after THA for 2 commonly used PROs, the Hip dysfunction and Osteoarthritis Outcome Score (HOOS) and the EQ-5D. I also aimed at estimating PASS for subgroups of age, sex and diagnoses.

METHODOLOGICAL CONSIDERATIONS

How to get the patients perspective

PROs reveal the patients perspective and the patient perspective is most important when quantifying and measuring pain, physical function and quality of life. But how should one best get patients to answer questionnaires? Response rate can vary considerably depending on patient group. The high response rate achieved in our study I and study III is however not only dependent on the patient group. I used several strategies to achieve this; I used relatively short PROs (maximum 2 A4 pages) with 6-19 items, had follow-up contact and provided a second copy of the PROs at follow up, mentioned an 'obligation' to respond (the results can lead to an improved treatment regimen for THA patients), used personalized PROs (patients name and identification number on the PRO), assured confidentiality and had a university sponsorship, as it is known that these factors contribute to a higher response rate (91). In study IV I printed copies of handwritten signatures in colored ink, to further personalize the patient correspondence (91). I also enclosed a return addressed envelope with a stamp (92). Despite the efforts only 73% of patients accepted participation in study IV. This may be explained by that the patients in this study received study invitation and information about the procedure close in time, which may have been a bit much information to process for many of the patients. In study IV there were 6 additional A4 pages of questions regarding patient characteristics besides the two PROs included, and the lengthier questionnaire could in part explain the lower percentage of participating patients (91).

Another important aspect of getting the patients perspective is the readability of PROs and correspondence. The text has to be easy to read and to understand for the patients (93). Choosing everyday language and avoiding medical terms is essential, and an important part of PRO development and PRO translation. I kept the included PRO's lay-out as close to the original as possible as not to change the measurement properties, with minimal

layout adjustments to optimize AFP readability. In the patient correspondence, I aimed at optimizing the lay-out, font type and point size to get the best possible readability for an elderly THA population. The peer-reviewed literature on readability and reading speed of different font types and point sizes, are sparse (94;95). I therefore consulted typographers and educationalists, and got the following advices; 1) what font is best, is dependent on media. 2) The correct point size is dependent on font. 3) Always avoid text in capital letters. 4) A sans-serif font like Verdana in point size 13-14 may be the best for paper printing, and therefore this was used in the correspondence and the patient information. The low proportion of items missing in study I and study II may, at least partly, be contributed to an acceptable readability of PROs and correspondence.

Feasibility

Several aspects of a PRO are important; there should be a published peer-reviewed development process, and preferentially several publications on usage in research and relevant clinical settings. As any other measure or measurement system, the different PROs have different measurement properties. The measurement properties of a PRO have often been called psychometric properties (or clinimetric properties), depending on the underlying theories or focus, but now a consensus is emerging (3) which may retire these older labels. Measurement properties often used, besides the ones in the 'List of terms and definitions', is measurement error (the systematic and random error of a patient's score that is not attributed to true changes in the construct to be measured) and responsiveness (the ability of an HR-PRO instrument to detect change over time in the construct to be measured) (3). To be able to choose the best PRO for a specific context, information on measurement properties has to be available. In addition to published development process, and to the measurement properties, the feasibility of using a PRO in a certain context is also important. The response rate, floor and ceiling effects, missing items, and need for manual validation in a specific context can be included in the definition of feasibility. To ensure generalizability and to minimize selection bias, a high response rate (of minimum 80% (93)) is usually considered to be adequate and sufficiently representative of the sample studied. To be able to measure deterioration and improvement of PROs, the floor and ceiling effects should generally be less than 15% (64). In postoperative THA patients, higher ceiling effects and lower floor effects can be expected, and a 15% ceiling effect might be too restrictive a criterion. This will be discussed further in the discussion section. A percentage of missing items of more than 5% (64) will lessen the validity of PRO data. If more than 5% of the scanned PROs are requiring manual validation (64) it is an important indirect indication of the patient's (lack of) general ability to correctly fill in the PRO, and also provide information about the workload of the manual validation required. The complexity of a PRO or the lack of comprehensiveness can therefore have an influence on the proportion of items missing as already mentioned, but also on the response rate, and the proportion of items requiring manual validation.

Besides the measurement properties of PROs, many other factors are important and ought to be considered when introducing a PRO into a registry setting. The PROs have to be present in the target language, and if not, translation, cross-cultural adaptation and validation are warranted. The feasibility has to be adequate if the data quality is to be acceptable, and achieving a good response rate is paramount (91). Some patient groups do not respond adequately to an Internet-based application for collecting PROs (7), and paper format questionnaires may have to be used. In this case the entire data collection systems should be examined with respect to data quality, especially when using newer techniques like AFP.

How to administer the PRO

Whether postal administration or internet-based administration is preferable, is dependent on patient population and setting; postal administration may have less desirability bias (93), but it may also be more challenging to get adequate response rates. Missing items and delay from late returned PROs, can also pose a problem. Internet-based administration may be cheaper, may have reduced erroneous responses due to no entry errors, but a risk of web-browser incompatibility, and low response rate if considered 'spam' by patients. Some patient groups are known to respond inadequately to an Internet-based application (7;96). Validation is a very complex matter if data is entered directly in an Internet-based application, since no other source of information exists to verify correctness of the data. The validity of Internet-based applications warrants further research as age and subgroup differences potentially may result in information bias.

PROs included in this thesis

The OHS (97) is an intervention- and site-specific outcome measure and this 12-item questionnaire is designed to assess functional ability, daily activities and pain, to get the THA patient's perspective. Items are answered by ticking a box on a five-point Likert scale and the raw scores are added to obtain a sum score (originally between 12 (worst) and 60 (best)), due to new recommendations the sum score should range between 0 and 48 with higher scores being better (98;99). OHS is reported to have an adequate reliability; a good internal consistency with a Cronbach's alpha of 0.84 (preoperatively) and 0.89 (six-month follow-up), and an intraclass correlation of 0.94 for the pre-operative data. Concerning construct validity, the OHS has been reported to correlate moderately with Charnley scores, and a significant agreement between OHS and the relevant scales of the SF-36 and the AIMS has been reported. OHS has been reported to have an acceptable sensitivity to change with effect sizes larger for OHS than for any of the scales of the SF-36 or the AIMS, indicating that the OHS may be particularly sensitive to improvements obtained by THA (99). The OHS has been translated into different languages and used in several clinical studies and in THA registry settings; it has been reported to be consistent, reliable, valid and sensitive to clinical change following THA (100-107). OHS cut-points associated with patient satisfaction with post-surgical outcomes have also been estimated (108). OHS have been mapped to the EQ-5D Index and a 0.02 point change in the EQ-5D Index was equivalent to a 1 point change in the OHS, where 42% of the variance was explained by the linear regression model (109). Academic and clinical use of OHS is free of charge. A license for the study and translation was obtained from Isis Innovation (http://www.isis-innovation.com/).

HOOS (110), is a hip-specific outcome measure and was constructed by adding items considered important by patients (concerning pain, symptoms, sport and recreation, function and hip-related quality of life) to the WOMAC (111) to get improved validity for those with less severe disease or higher demands of physical function. The HOOS includes 5 subscales: Pain, Other Symptoms, Function in Daily Living, Function in Sport and Recreation and Hip-related Quality of Life. HOOS Physical Function Short form (HOOS PS) is a 5-item short version derived

from the two HOOS subscales: Function in Daily Living and Sport and Recreation Function, and was developed using Rasch analysis (112) by using data from samples representing a wide spectrum of OA severity (113). The HOOS PS has been validated for THA (114). I used three different HOOS subscales in our studies; HOOS Pain, HOOS PS and HOOS Hip-related Quality of Life (QoL) to measure pain, physical function including daily activities and more strenuous physical activities, and hip-related quality of life. The sum scores of the subscales range between 0 and 100 with higher scores being better. HOOS does not require any license and is free of charge, even to the medical industry. User guide and a scoring manual are available at http://www.koos.nu/index.html.

EQ-5D (115;116) is a generic health outcome measure, and is applicable to a wide range of health conditions and treatments by identifying 243 possible health states. EQ-5D can be used for economic evaluation of health care, and is designed to complement other 'quality of life'-measures, or disease-specific outcome measures. Patients describe their own health state on 5 dimensions: mobility, self-care, usual activities, pain/discomfort and anxiety/ depression, and one of three levels of severity is chosen for each dimension (in the version used): no problems, some/ moderate problems or extreme problems. Patients also value their current state of health on a thermometer scale from 0 ('worst imaginable') to 100 ('best imaginable'), and the EQ-5D therefore generates two overall values for the quality of life, one from the patient's perspective (the EQ-VAS; 'Current state of health') and the other from a societal perspective, the EQ-5D Index (a health profile that can be made into a global health index with a weighted total value for health related quality of life), which represent the patients description of their own health and how this health state is perceived by the general population. I used a Danish tariff (117) based on time-trade-off (118) when computing the index to adjust for cultural differences in response pattern, and the Index ranged from -0.624 (worst) to 1.000 (best). In 2001 the EQ-5D was validated for THA patients (119), and in 2009 for rheumatoid arthritis (RA) patients (120). The EQ-5D is currently used in the Swedish Hip arthroplasty Registry (7). Academic and clinical use of EQ-5D is has been free of charge if patient numbers are less than 5,000. Where patient numbers exceeded 5,000, the EuroQol Group would negotiate with users to collaborate and share data. However, the policy for routine use of EQ-5D is currently under revision. License to the study was obtained from the EuroQol Group (http://www.euroqol.org/).

SF-12 is a generic health outcome measure (121), which has been validated on OA patients (122). It consists of 12 items derived from the 36-item score, SF-36 (123). The SF-12 gives two summary scores; Physical Component Summary (PCS) and Mental Component Summary (MCS), ranging from 0 to 100 with higher scores being better. The sum scores are calculated in the special QualityMetric Incorporated´s scoring software by computation with a standardized scoring algorithm developed to get a mean of 50 and a standard deviation (SD) of 10 in the US 1998 general population value set. The fees associated with using SF-12 were altogether 1,569.90 USD (administrative fee, survey reference kit and scoring software). PCS and MCS were treated as one variable in the analyses, since they are derived from the same items but with different weighting, due to dependence. License to the study was obtained from the Medical Outcomes Trust Health Assessment Lab and Quality Metric Incorporated (http://www.sf-36.org/).

Selection of PROs for the studies

In study I-III four different PROs were included; EQ-5D, SF-12, HOOS and OHS. These PROs were chosen after a literature search, and the aim was to find two general health PROs and two hip-specific PROs, who all were relatively short (max 2 A4 pages), all commonly used in the orthopedic field and all having documented adequate measurement properties. I chose only to include outcome measures reported by patients and not surgeon reported outcomes, as the main importance was the patient perspective, and surgeons tend to rate the patients outcome different than the patients themselves (47). Patients in study I-III each received one general health PRO and one hip-specific PRO in four groups receiving different PRO combinations, and I cannot completely rule out that the combinations of the PROs affected the answers. I also cannot rule out that the different number of items in the included PROs affected the results. Since all PROs had a similar length (2 A4 pages) it is unlikely that the different number of items in the PROs gravely affected the results.

In study I, I concluded that the HOOS, the OHS, the SF-12, and the EQ-5D were all appropriate PROs for administration in a hip registry, but in study IV, only two PROs were included; HOOS and EQ-5D. I wanted to include one general health PRO and one hip-specific PRO and chose only to include two PROs to reduce the patient burden (124). The differences found between the PROs in study I were minor, and HOOS was chosen over OHS because of the subgroup division in HOOS. By using HOOS, three outcomes were collected; pain, physical function and quality of life. Using OHS would render only one sum score, linked to the quite unspecific domain "hip problems". EQ-5D were chosen over SF-12 due to easier license requirements, lower fees associated with usage, no requirements of a specific scoring software and also because of the successful inclusion of EQ-5D in the Swedish Hip Arthroplasty Registry (7). HOOS includes WOMAC in its complete and original format, and WOMAC scores can be calculated. A review by Ahmad et al. recommend to use a combination of OHS and WOMAC (125).

Importance of registration; PROs in the national joint registries

The shift towards a more patient-centered perspective and an increase in the use of PROs (52), has also been reflected in the measurement practices of the regional and national registries, where more and more joint registries, for example the Swedish Hip Arthroplasty Register, the New Zealand Joint Registry, the National Joint Registry for England and Wales, the California Joint Replacement Registry, the Winnipeg Regional Health Authority Joint Replacement Registry and the Center for Education and Research on Therapeutics Registry, are collecting PROs (57;58).

Translation

There are good reasons to translate a PRO instead of making a new; first there are many high quality PRO already available. Second, it requires much time and effort making an adequate PRO. Third, a translation makes it possible to compare results internationally. Several guidelines exist (126;127), and lot of effort has been made to established a best-practice methodology for the translation and cross-cultural adaptation process (66). Most guidelines have the steps shown in Table 2, in common. In study III, I used a strict methodology for translation and cross-cultural adaptation (66) and I am confident that I have found the best possible Danish wording, while attaining the conceptual agreement for the Danish language version of OHS. There were only minor discrepancies concerning wording and understanding in the translation process, probably due to the relatively small

cultural difference between England and Denmark. In item 6 (Walking time before severe pain) instead of the original option 4, 'around the house only', I chose to focus on walking distance ('only very short distances'). The Danish option 4 ('only very short distances') implies that the person is housebound, especially since this option is situated between the options '5 to 15 minutes' and 'Not at all/pain severe on walking'. I chose to focus on walking distance for this option for item 6, because I am not sure that the UK and the Danish concept of 'housebound' is equivalent, or equivalently dependent on walking ability, due to the differences in the size and the number of floors in homes in Denmark compared with England. Item 3 (Trouble with transport) is a complex question and consists of three different questions; 'Have you had any trouble getting in a car because of your hip?', 'Have you had any trouble getting out of a car because of your hip?' and 'Have you had any trouble using public transport because of your hip?' The testing showed that some patients were unsure of how to answer, if they answered yes to only one or two of these questions. To resolve this problem, I added Danish written instructions to the OHS, as an addendum (Paper III, Supplementary Material).

Table 2. Translation of PROs

Translation	
Step	**Important aspects**
Forward translation	**Conceptual rather than literal translations, bilingual translators, mother tongue of the target culture, simple, clear and concise language, avoid the use of any jargon, consider issues of gender and age applicability, avoid terms considered offensive**
Expert panel discussion	**Bilingual expert panel, multidisciplinary group, identify and resolve inadequate expressions/concepts, identify and resolve discrepancies between versions**
Back-translation	**Independent translators, mother tongue language of original PRO, no knowledge of the original PRO, conceptual and cultural equivalence, discrepancies should be discussed, forward translation/ back translation as many times as needed until a satisfactory version is reached**
Pre-testing and cognitive interviewing	**Pre-test respondents representative of patient group, 10 minimum, represent males and females, from all age groups (18 years of age and older), pre-test respondents systematically debriefed**
Final version and documentation	**Final version result of all the iterations described above, all the cultural adaptation procedures should be documented**

The clinical relevance of PROs; MCII and PASS

In parallel with the shift towards a more patient-centered perspective and the change in focus from traditional clinical outcomes to PROs (52;58), there has been an increased interest in how to best interpret PRO results (68). This is easy to understand since the interpretation of PRO change scores and postoperative PRO scores can be very problematic (128). What is the clinically meaningful interpretation of a postoperative HOOS Pain score of 81? What does it mean if a patient has a change score after the operation in EQ-VAS of 21?

MCII and PASS can help answering these questions. The MCII is the minimal difference representing a clinically important difference in the patient's perspective, in the direction of improvement (129). The PASS reflects the overall health state at which patients consider themselves to be feeling well (130). There is a lack of these kind of cut-point estimates in the musculoskeletal literature (131) and since MCII and PASS estimates are not constant for a single PRO, but rather dependent on the context in which the PRO is used, continued estimations are required to step-by-step contribute to our understanding of how to interpret change in and absolute PRO scores following orthopedic procedures.

Different estimation methods exist for estimating MCII (132-134) and PASS (135-141). The main division of estimation types is the anchor-based methods and the distribution-based methods (132). Since the focus of study IV was the patients' perspective, only anchor based methods were used. MCII and PASS estimates were calculated by multiple approaches, further outlined in the section concerning statistical methods.

The quality of a HR-PRO is dependent on the documented validity, reliability, responsiveness and interpretability. MCII and PASS estimates contribute to the interpretability of the PROs. To further enhance the interpretability, distribution based reliability measures for change scores have been calculated. These measures can help validate anchor based MCIIs, as they give information on the possibility of detecting the patient reported MCII, with an adequate precision.

Distribution–based measures

Distribution-based methods for MCII estimation is without anchoring, and therefore without information regarding the patient perspective, but can be used as an approximation where no other MCII has been estimated. In addition to this, the different distribution-based methods can be used to examine the precision and variation of anchor-based MCII, as the distribution-based measures are based on statistical properties of the PROs.

The SD of change has been used as a distribution-based reliability measure, and it has been suggested that ½ SD can be used as an approximate MCII (142). Limits of agreement (LOA) gives information on how random variation influence observations, by calculating 1.96 standard deviations of the mean bias. Using the Bland-Altman method in non-independent data have been criticized as this approach is not suitable for repeated measures data, but it may however be used to explore the data (143). The LOA is expressed in the units of measurement and indicates the size of the measurement error. A Bland–Altman plot shows the difference of each point, the mean difference, and the limits of agreement on the vertical axis and the average of the two ratings on the horizontal axis. Thus the Bland–Altman plot demonstrates both the overall degree of agreement and whether the agreement is related to the underlying value of the item and offers a graphic visualization of the change in preoperative- to postoperative status and the test-retest item- and sum score agreement.

The standard error of the mean, calculated as SD change / √n, represent the standard deviation of the error in the sample mean relative to the true mean. The minimal detectable change (MDC) (132;144;145) or smallest detectable change (146), calculated as 1.96 x √2 x standard error of measurement (SEM), describe which changes that fall outside the measurement error of the PROs. The effect size (ES), (ES = Δ / SD baseline) describes the sensitivity of PROs for detecting clinical change (133;134;147-150). ES of 0.2–0.5 can be regarded as small, 0.5–0.8 as moderate and ES above 0.8 as large (148). The standardized response means (SRM) (134;150-153), (SRM = Δ / SD change), is similar to ES, but is

calculated by dividing the mean change by the standard deviation of the change scores (not the standard deviation of the baseline scores). SRM of 0.2–0.5 can be regarded as small, 0.5–0.8 as moderate and ES above 0.8 as large (153). ES and SRM (and also the responsiveness index) are methods based on sample variation.

The SEM, calculated as SD baseline √ (1-reliability), is an often used measure (132;133;144;154-158). A test-retest reliability (1) of 0.89 for HOOS Pain (159), 0.86 for HOOS-PS (160), 0.78 for HOOS QoL (159), 0.82 for EQ-5D Index (7) and 0.83 for EQ-VAS (7) was used for calculating SEM. The reliability change index (RCI) (156;161-163), (RCI = Δ / √2 x SEM), is closely related to the MDC (i.e., 1.96 x √2 x SEM), and describes the standard error of the measurement difference. Both SEM and RCI are methods based on measurement precision.

Examples of other distribution-based reliability measure not included in the thesis are the responsiveness index (calculated from the distribution as the ratio of the mean change in score after treatment to the variability in stable subjects) (164), and the relative efficiency (the ratio of the square of the t-statistic of a comparator PRO over the square of the t-statistic of the reference PRO) (134;165).

Anchors – getting the patients interpretation of PRO scores

To be able to estimate MCII and PASS based not solely on the distribution, but based on the patients perspective, anchor items are imperative. An anchor item is often a retrospective global transition question, or a clinical anchor (132), but also an absolute change anchor can be used (166). The anchor item establishes a connection between the PRO change scores or the postoperative PRO scores and patients' health situation. In study IV a self-reported hip-specific anchor question was used for MCII estimation, a self-reported hip-specific anchor question was used for PASS estimation and one self-reported general health anchor question (167) was used for MCII and PASS estimation. These anchor questions are used in 'Questionnaire for patients who have had hip surgery' from The Royal College of Surgeons of England (168) and have been used and studied in large populations (169;170). The anchors describe changes in hip symptoms from preoperatively to postoperatively, postoperative hip symptoms states, general health changes and general postoperative symptoms states, respectively.

Information bias

In addition to the usual source of biases (see the 'Strengths and limitations' section), PROs are known to be prone to information bias, heuristics and cognitive biases (171). I had several strategies to minimize this (93). I minimized information bias by using well documented questionnaires, with relevant questions. I had a patient group who wanted to 'share their story', and ensured no 'item over-kill'; Only 6 (EQ-5D), 12 (OHS and SF12) and 19 (HOOS-Pain-PS-QoL) items. I had relevant information in the invitation letter. There were few, if any, embarrassing items (e.g. sex life) and few dichotomous 'Yes/ No' questions (in total 4 items in SF12). I used PROs with carefully chosen wording and less positive or negative connotations. The answer categories were relevant and 37 (of total 49) items have 5-6 possible answers (5-9 possible answers often considered optimal (93)). No evident external interests were present. Recall bias is known to be a problem for retrospective items (132), and in study IV, I used both a retrospective anchor and a change anchor for MCII estimation, to account for this.

Missing items

For the different PROs, I handled missing items in accordance with the directions set out in the specific manual for each PRO in question; for EQ-5D I used no imputing of missing values (172), for SF-12 I used QualityMetric Incorporated´s scoring software (version 2.0 and 4.0) which includes an MDE algorithm that enables scoring of PCS and MCS with missing item responses and I used the missing data estimation method; maximum data recovery (the exact procedure is not described (173)), to find percentage of discarded PRO subscales. In the other analyses, I used manual coding with no imputing of missing values. For HOOS, one or two missing values were substituted with the average value for that subscale. If more than two items were omitted, the response was considered invalid and no subscale score was calculated (174). For OHS, if one or two items were unanswered, I entered the mean value representing all of the patients other item responses, to fill the gaps, but if more than two items were unanswered the overall score for that patient was not calculated (99).

STATISTICAL METHODS

Descriptive statistics

Categorical variables are presented as frequencies and proportions. Continuous variables are presented as means and 95% confidence intervals (CI) or standard errors (SE), or median and ranges.

In paper I the response rate, floor and ceiling effects, missing items, and the need for manual validation were calculated as proportions with 95% CIs. The defined cut-points for all 5 criteria in order to identify PROs that were feasible for use in registry settings were: overall response rate over 80%, floor and ceiling effects less than 15%, a proportion of items missing of less than 5%, and a proportion of items needing manual validation of less than 5%.

In paper II the error proportions were calculated as proportion of errors per 1,000 data field with binomial exact 95% CI (STATA procedure 'cii'). Validation of the AFP in relation to person ID, was done in comparison with the original sample of all patients (n=5,777), with STATA 'assert' command.

In paper III the response rate, floor and ceiling effects, and missing items were calculated as proportions with 95% CIs. For test-retest, I used the STATA 'sample' command to draw random samples of the original cohort from the Danish Hip Registry.

In paper VI, I calculated the proportions (percent) of patients reporting different response categories to the anchor questions and the corresponding PRO change scores and postoperative PRO scores. The absolute scores of the different HOOS subscales, EQ-5D Index and EQ-VAS were calculated preoperatively and postoperatively for each individual patient, as well as change scores from pre- to postoperatively. I also calculated mean (95% CI) preoperative and postoperative PRO scores and mean change scores (95% CI) for the entire study population. I estimated PASS (95% CI) for subgroups of different sex, diagnoses and age. Due to small subgroups MCII were not estimated at subgroups level, but I calculated mean (95% CI) PRO change scores for the different subgroups included.

Comparing the mean or proportions

I used chi-square test (two nominal variables), Student's t-tests (nominal and interval variables) and Wilcoxon-Mann-Whitney test (ordinal or interval variables) to compare responder and non-responder characteristics, and to otherwise compare proportions.

In paper IV, Welch's t-test or a W test (175;176), both allowing for unequal variances across groups, was used for comparing means between subgroups.

In paper II, I studied the error proportion overall and for each of the four different questionnaires, and also for each individual patient. This was tabulated in subgroups by sex and age groups (<60 years, and >60 years) with binomial CIs. Due to the prespecified and low number of tests, I saw no reason to adjust the p-level by multicomparison principles. Throughout this thesis a two-tailed probability value less than 0.05 is considered significant.

Regression models

In paper I, logistic regression was used to compare overall feasibility criteria between different PROs, adjusting for age (< 50, 50–70, and > 70 years), sex, primary hip diagnosis (idiopathic OA, inflammatory arthritis, childhood diseases, high-impact injuries, and low-impact fractures) and prosthesis type (uncemented, cemented, or hybrid). Odds ratios (OR) with 95% CIs were calculated.

I studied the abilities of different PRO subscales to discriminate between age and sex groups, diagnostic groups, and prosthesis types using analysis of variance. The hypothetical number of subjects needed to find the significant difference in mean value of a PRO between groups (assuming a significance level of 5% and a power of 85% to detect differences between the actual groups) was estimated for each PRO subscale with sample-size calculations or with power calculations and simulated ANOVA F tests, depending on the number of groups.

Correlations

In paper III the construct validity was tested by comparing the Spearman's correlation coefficients. Internal consistency was determined by calculating Cronbach's Alpha. Intraclass correlation (ICC) was calculated as ICC agreement[2,1] (64) and ICC consistency [3,1] (177;178) with STATA 'icc23' command (two-way random effects model). Bland and Altman's limits of agreement were calculated by STATA 'concord' command and Bland-Altman plots were made using STATA 'batplot' command.

In paper IV the correlation between the anchors and the PRO and PRO subscales were tested with Spearman's correlation coefficients. Cohen's guidelines for interpreting the magnitude of correlation coefficients (r = 0.1 (small), r = 0.3 (moderate), and r =0.5 (large)) were used (148).

MCII and PASS

In paper IV, the MCII and PASS estimates were calculated by multiple approaches: the mean change or mean score approach (135;137-140;179), the 75th percentile approach (135;137-140;166), the 75th percentile approach using tertiles (lowest-, middle-, and highest subscale scores) of the preoperative PRO scores (180;181), and the following receiver operating characteristic (ROC) curves methods; the 80% specificity rule (137;181;182), the cut-point corresponding to the smallest residual sum of sensitivity and specificity (135;137;140;141) and the cut-point corresponding to a 45 degree tangent line intersection (equivalent to the point at which the sensitivity and specificity are closest together) (141). The mean change approach and the mean score approach were used as the primary approaches for MCII and PASS, respectively. 95% CI for cut-points were estimated by non-parametric bootstrap (182;183) using 2000 replications, since some groups were small (n<30) in the tertiles estimations. The area under the curve (AUC) with 95% CI was calculated for all three methods using ROC curves.

Factor analyses

Exploratory factor analyses by principal component analysis with polychoric correlations were conducted for all included multi item PROs or PRO subscales in study I. Threshold for factor loadings were set at 0.5 (184). Confirmatory factor analysis is most often used to assess structural validity, but no STATA module for confirmatory factor analyses with the correct statistical assumptions could be found.

Differential item functioning

Analyses of DIF were performed for OHS on the following groups; time since operation (1-2 years, 5-6 years, 10-11 years), age group (<50, 50-70, >70), and sex. Significance level 0.05/12 was used to correct for multiple testing (Bonferroni correction). A cut-point of minimum 10% change in effect size (beta) as a criterion for clinically relevant DIF was used.

Software

The R software Version 3.0.1 (The R Foundation for Statistical Computing, Vienna, Austria) with "lordif" package was used for differential item functioning. The STATA software Version 10.1 and 11.0 (StataCorp LP, Texas, USA) was used for all other statistical analyses.

SUMMARY OF PAPERS

PAPER I

In this study I compared the feasibility of the four PROs examined; EQ-5D, SF-12, HOOS and OHS. I tested response rates, floor and ceiling effects, missing items, and need for manual validation of forms. I also calculated the number of patients needed for each PRO to discriminate between subgroups of age, sex, diagnosis, and prosthesis type in a hypothetical repeat study.

Paper I describes a sample of 5,777 patients (all patients over 18 years of age) registered in DHR with a primary THA and who underwent surgery 1–2, 5–6, and 10–11 years prior to the study. These current analyses include 5,747 THA patients registered in the DHR.

Results

Response rate

All PROs fulfilled our criteria of an overall response rate of over 80%. Multiple regression analyses adjusted for age, sex, diagnosis, and type of prosthesis showed no overall difference in the response rate for HOOS and OHS (adjusted OR = 0.90, CI: 0.78–1.04). For the generic PROs the overall adjusted OR for response rate was 1.12 (CI: 0.97–1.30). Separate multivariate analyses of differences in response rate for disease-specific PROs and generic PROs showed similar results for females and for different age groups. However, males who had received the EQ-5D responded more often than males who had received the SF-12 (adjusted OR = 1.4, CI: 1.1–1.8).

Floor and ceiling effects

All PROs fulfilled our criteria of a floor effect of less than 15%; the floor effect was 0.5% or less for the disease-specific PROs ($p < 0.001$) and less than 0.3% for the generic PROs ($p = 0.03$). However, neither the HOOS nor the OHS fulfilled our criteria of a ceiling effect of less than 15%. SF-12 PCS and MCS and the EQ-

VAS fulfilled our criteria of a ceiling effect of less than 15%, while the EQ-5D Index had a high ceiling effect of 45.8% ($p < 0.001$).

Missing items and discarded subscales
All PROs fulfilled our criteria of a proportion of items missing of less than 5%. The percentage of discarded PRO subscales, where a score could not be calculated due to too many missing items, was between 1.2% and 3.0% for disease-specific PROs ($p < 0.001$) and between 2.3% and 5.5% for generic PROs ($p < 0.001$). With multivariate analysis, I found a significantly higher risk of discarded PROs for female patients with HOOS Pain, HOOS PS, and HOOS Qol compared to patients with OHS. For the generic PROs, the EQ-5D Index and EQ-VAS had a higher risk of discarded questionnaires than SF-12 PCS/ MCS; adjusted OR for EQ-5D Index was 1.4 (CI: 1.0–2.1) and for EQ-VAS it was 2.6 (IC: 1.9–3.6).

Manual validation
All PROs fulfilled our criteria of a proportion of items requiring manual validation of less than 5%. However, the proportion of questionnaires requiring manual validation exceeded 7% for all PROs. For the generic PROs, 7.7% of the items in the SF-12 questionnaires required manual validation as compared to 21.8% in the EQ-5D questionnaires ($p < 0.001$).

Discriminative ability
Group sizes from 51 to 1,566, depending on descriptive factors and choice of PRO, were needed for subgroup analysis. OHS had the best discriminative ability—described by the hypothetical number of subjects needed to discriminate between groups in relation to gender (298 patients per group were needed to find a statistically significant difference in mean sum score). SF-12 PCS had the best discriminative ability in relation to diagnosis (51 patients per group were needed). EQ-VAS had the best discriminative ability regarding both age (where 270 patients per group were needed) and prostheses type (where 207 patients per group were needed).

PAPER II
In this study I assessed the quality of AFP and validated an up-to-date AFP system, by comparing paper-based and scanned patient-reported outcome forms with single and double manually entered data.

Paper II describes 200 patients randomly selected from the patient cohort of Paper I. The analyses included 200 THA patients, 398 PROs, 4,875 items and 21,887 data fields.

Results
ICR
There was no statistically significant difference between double-key entering (error proportion per 1,000 fields = 3.367 (95% CI: 0.085–18.616)) and single-key entering (error proportion per 1,000 fields = 6.734 (95% CI: 0.817–24.113), ($p = 0.565$)), no statistical difference between AFP (error proportion per 1,000 fields = 10.101 (95% CI: 2.088–29.234)) and double-key entering ($p = 0.319$), nor any statistical difference between AFP and single-key entering ($p = 0.656$).

OMR
AFP (error proportion per 1,000 fields = 0.046 (95% CI: 0.001–0.258)) performed better than single-key entering (error proportion per 1,000 fields = 0.370 (95% CI: 0.160–0.729), ($p = 0.020$)), double-key entering (error proportion per 1,000 fields = 0.046 (95% CI: 0.001–0.258)) performed better than single-key entering ($p = 0.020$), and AFP and double-key entering performed equally ($p = 1.000$).

PROs, gender and age
I found no difference in performance for the different questionnaires with the AFP in OMR ($p = 0.609$), with double-key entering ($p = 0.644$), or single-key entering ($p = 0.148$). Concerning gender, I found no statistical differences for ICR ($p = 0.304$, $p = 0.239$, $p = 0.095$), or OMR ($p = 0.409$, $p = 0.409$, $p = 0.371$). Similarly, there were no differences concerning age for ICR ($p = 0.520$, $p = 0.711$, $p = 0.711$), or OMR ($p = 0.687$, $p = 0.687$, $p = 0.904$).

PAPER III
In this study I translated and cross-culturally adapted the original OHS into Danish and validated the Danish language version by testing the measurement properties.

Paper III is a secondary analysis of data from Paper I, including a subgroup of all patients between the ages of 30 and 80 years who had previously answered the OHS and 215 patients who had previously answered the HOOS, giving a total of 2,278 patients for this study. For test-retest validation, 212 patients received the OHS twice within two weeks.

Results
Translation and Cross-Cultural Adaptation
The translation process revealed minor discrepancies in wording and understanding for items 1 (Usual level of hip pain), 8 (Pain on standing up from sitting), 9 (Limping when walking), 11 (Work interference due to pain), 12 (Pain in bed at night) and option 4 in item 6 (Walking time before severe pain), so these were rephrased in the translation process. Some patients had problems with item 3 (Trouble with transport), which I resolved by adding a written instruction for the questionnaire.

Psychometric properties
The OHS had a response rate of 87.4%, no floor effect and 19.9 % ceiling effect in our postoperative patients, and one per cent of patients had too many items missing to calculate a sum score. The frequency distribution of the scores was negatively skewed, with a skew value of -1.39.

Regarding construct validity, OHS showed the highest correlations with the HOOS Pain, HOOS PS and HOOS QoL; the pain/ discomfort domain, mobility, current state of health and the usual activities domain from the EQ-5D; and the body pain domain from the SF-12 (rho = +/- 0.51 to 0.62). The OHS showed the lowest correlations with the anxiety/depression and self-care domains of the EQ-5D; and the mental component score, vitality and social functioning domains from SF-12 (rho = +/- 0.32 to 0.46). SF-12 general health, body pain domain and physical component score had a correlation of 0.38 to 0.49. Thus 12 of the 15 predefined hypotheses about the strength of correlation were confirmed.

The test-retest reliability of the OHS sum score was established with an ICC of 0.96 (95% CI: 0.94–0.97), and limits of agreement was -0.05 (95% CI: -4.67–4.58). For internal consistency, the overall Cronbach´s alpha was 0.99, and the average inter-item correlation was 0.88.

PAPER IV
In this study I estimated MCII and PASS for HOOS subscales and for the EQ-5D in THA patients.

Paper IV describes all patients over 18 years receiving a THA in one of 16 orthopedic departments in Denmark ('Odense Universitetshospital', 'Middelfart Sygehus', 'Vejle Sygehus', 'Nordsjællands Hospital Hillerød-Hørsholm', 'Privathospitalet Hamlet', 'Sygehus Sønderjylland-Sønderborg', 'Frederiksberg Hospital', 'Klinik Aalborg, Aalborg Sygehus Syd', 'Herlev Hospital', 'Gentofte Hospital', 'Erichsens Privathospital', 'Friklinik Frederikshavn', 'Nykøbing Falster Sygehus', 'Regionshospitalet Viborg', 'Holbæk Sygehus' and 'Næstved Sygehus'), from 01.03.10 to 01.03.11, and who accepted study participation, giving a total of 1,335 patients for this study. The patients were followed from the preoperative assessment to one year postoperative.

Results

MCII

MCII cut-points for HOOS based on the hip-specific anchor question 'Overall, how are the problems now in the hip on which you had surgery, compared to before your operation?' were 24 (95% CI: 20-28) for HOOS Pain, 23 (95% CI: 19-28) for HOOS PS and 17 (95% CI: 12-22) for HOOS QoL. The estimated MCII cut-points for EQ-5D Index and EQ-VAS based on a general health anchor were 0.31 (95% CI: 0.29-0.34), and 23 (95% CI: 21-25), respectively. MCII estimates were dependent on baseline score for all PROs, since lower tertiles corresponded to higher MCII estimates.

PASS

PASS cut-points for the HOOS subscales when responding 'Excellent', 'Very good' or 'Good' to the question 'How would you describe the results of your operation?' were 91 (95% CI: 91-92) for HOOS Pain, 88 (95% CI: 87-89) for HOOS PS and 83 (95% CI: 82-85) for HOOS QoL. The cut-points representing PASS when reporting 1 step better general health postoperatively compared to preoperatively were 0.92 (95% CI: 0.91-0.92) for EQ-5D Index, and 85 (95% CI: 84-86) for the EQ-VAS. PASS estimates were independent of preoperative score as shown by identical PASS cut-points for the different tertiles of baseline scores.

Males had better PASS estimates than females ($p \leq 0.04$), and idiopathic OA patients had better PASS estimates for HOOS QoL and EQ-5D Index than other patients ($p \leq 0.008$) and patients over 70 years had lower PASS estimates than younger patients for HOOS Pain, HOOS-PS and EQ-VAS ($p \leq 0.03$).

DISCUSSION

Measurement properties

PROs are measurement instruments and it is therefore important to have knowledge about the measurement properties of the different PROs. The PRO results have to be valid, reliable and responsive, the PROs have to be feasible to use and the results have to be interpretable. The COSMIN study has tried to unify the definitions and the taxonomy of relationships of measurement properties, and lists many of the different measurement properties considered most important (3).

Study III examined measurement properties of the Danish language version of OHS. I consider the other PROs used in the studies (HOOS, SF-12 and EQ-5D) to be sufficiently validated, and have therefore not done a full validation of these. In study I, mean PRO scores for the total population have been reported (Paper I, Table 2), now accompanied by median score, IQR and range (Table 3). In study III the convergent and divergent construct validity of the Danish OHS were assessed by hypothesis testing and found adequate with over 75% of the predefined hypotheses confirmed (64), which correspond well to other findings between the OHS and the HOOS (185). The OHS' correlation with SF-36 has also been found to be moderate to high for the physical function and bodily pain domains in postoperative patients (99;106). No studies of OHS' correlation with SF-12 could be found.

Table 3. Additional results: Median, Interquartile range and Range of PRO scores, from Study I

PROs	Median	Interquartile range (IQR)	Range
HOOS Pain	95	80-100	0-100
HOOS-PS	90	70-100	0-100
HOOS QoL	88	63-100	0-100
OHS	43	34-47	0-48
SF-12 PCS	43	33-45	12-63
SF-12 MCS	56	46-61	12-70
EQ-5D Index	0.84	0.72-1.00	-0.33-1.00
EQ-VAS	85	70-95	0-100

Concerning test-retest reliability, the ICC should be above 0.70 to be acceptable (64). The ICC of the different items in OHS ranged from 0.80 to 0.95, and the OHS sum score had a LOA of -0.05 (-4.67 to 4.58) and an ICC of 0.96 (0.94 to 0.97). This is higher than in the original OHS and in other language versions (99;105;107), and may be explained by the postoperative administration of the OHS in our study.

Table 4. Additional results: Variance components for OHS sum score, from Study III

Variance	Estimate (95% CI)
Between patients	64.4 (51.7-80.2)
Within patient	2.8 (2.2-3.5)

Table 5. Additional results: Distribution based measures of change in OHS, from Study III

n=166	OHS
Mean change score	0.05
Standard deviation (SDchange)	2.36
Standard error of measurement (SEM)	1.60
Effect size (ES)	0.01
Minimal detectable change (MDC)	4.45
Standardized response mean (SRM)	0.02
Standard error of the mean	0.19
Reliability change index (RCI)	0.02

Table 6. Additional results: Intra Class Correlation (ICC Consistency) of OHS, from Study III

Question	Content *1*	ICC (95% CI)
1	Usual level of hip pain	0.87 (0.83-0.90)
2	Trouble with washing and drying	0.85 (0.80-0.89)
3	Trouble with transport	0.80 (0.74-0.85)
4	Putting on socks/stockings/tights	0.85 (0.80-0.89)
5	Doing household shopping alone	0.95 (0.93-0.96)
6	Walking time before severe pain	0.79 (0.73-0.84)
7	Difficulty going up stairs	0.84 (0.79-0.88)
8	Pain on standing up from sitting	0.82 (0.77-0.87)
9	Limping when walking	0.81 (0.75-0.86)
10	Sudden, severe pain from hip	0.86 (0.82-0.90)
11	Work interference due to pain	0.85 (0.80-0.88)
12	Pain in bed at night	0.86 (0.81-0.89)
OHS sum score		0.96 (0.94-0.97)

1: The wording of each item reported in this table is in abridged form

The variance between the patients for OHS sum score was 64.4 (95% CI: 51.7-80.2), and the variance within the same patient was 2.8 (95% CI: 2.2-3.5) (Table 4). A systematic error variance could have a greater relative impact if the variance between the patients was lower. The MDC (4.45) was within the LOA of the OHS sum score (Table 5). ICC consistency of the different items in OHS ranged from 0.79 to 0.95, and the OHS sum score had an ICC consistency of 0.96 (0.94 to 0.97) (Table 6).

A Cronbach's alpha over 0.95 could be explained by a possible redundancy in one or more items (64), but seems to rise directly in line with the length of follow-up (Cronbach's alphas of 0.87 to 0.89 have been reported in preoperative patients (106;107), 0.89 at 6 months postoperative (102;186), and 0.93 to 0.92 at one to two years postoperative (102)). The very high internal consistency of the OHS found in Study III, with a Cronbach's alpha of 0.99, is almost certainly due to the long follow-up period, where patients are likely to have few or no symptoms giving a suboptimal timeframe to assess the Cronbach's alpha, and are therefore not due to item redundancy. The alpha would decrease to 0.89-0.96, if any (one) item was removed.

Responsiveness of HOOS and EQ-5D was assessed by hypothesis testing, 46% of the hypotheses were rejected, and the responsiveness was considered moderate (Table 7). The responsiveness of HOOS and EQ-5D has previously been assessed by calculation of SRM, and will be discussed further in the section concerning anchor–based and distribution–based measures.

Even though PROs have been increasingly studied in joint registry contexts in recent years (4;7), still many aspects of their use in this context warrants further examinations and the full potential of registry PRO usage is far from reached (58). Some examples of this is the lack of registry studies identifying inferior THA implants by the use of MCII and PASS for PROs, the lack of registry studies identifying inferior THA surgery approaches by the use of MCII and PASS for PROs, and the lack of registry studies identifying THA patients at risk by the use of MCII and PASS for PROs. Study I is to my knowledge the first feasibility study comparing commonly used hip-specific and generic PROs head-to-head in a hip registry setting, and Study III is the first translation, cross-cultural adaptation and validation study of a Danish language version of OHS, showing that feasibility studies and validation studies of Danish orthopedic PROs, are in its infancy. Validity, feasibility, response rate, ceiling effect, missing data, manual validation, factor analyses and differential item functioning will be discussed in the following sections.

Validity and feasibility

Validity and feasibility are two of several important measurements properties for measuring the quality of a PRO. Validity of a PRO can be defined as 'the degree to which a health related PRO instrument measures the construct(s) it purports to measure' (3). Besides a good validity, the PRO also has to be feasible to use in the intended context. I have defined an adequate feasibility in a registry setting as where the response rate is over 80%, the floor and ceiling effects are less than 15% (see also the discussion of ceiling effect below), the proportion of items missing is less than 5% and the need for manual validation of the scanned PROs is low, with a proportion of items needing manual validation of less than 5%, thus replacing an older and more general definition of feasibility as 'the average usable response rate for a questionnaire in a postal survey'(4). This older definition of feasibility takes into account the response rate, and combines this with the amount of missing items or the completion rate ('average usable response rate'). In this thesis it is argued that these two properties are not enough for measuring the feasibility of a PRO in a specific context. For example, if the ceiling effect of a PRO preoperatively was close to 100% in an intervention study, the PRO would clearly not be feasible to use in

Table 7. Additional results: Responsiveness of HOOS and EQ-5D, from Study IV

	Hypotheses *1*	Correlations *2*	Confirmed
1	**The correlation of change on HOOS Pain with EQ-5D item 4 (pain/discomfort) is at least 0.10 higher than the correlation of change on HOOS-PS with EQ-5D item 4**	**-0.52 vs -0.45**	**No**
2	**The correlation of change on HOOS Pain with EQ-5D item 4 (pain/discomfort) is at least 0.10 higher than the correlation of change on HOOS QoL with EQ-5D item 4**	**-0.52 vs -0.50**	**No**
3	**The correlation of change on HOOS-PS with EQ-5D item 1 (mobility) is at least 0.10 higher than the correlation of change on HOOS Pain with EQ-5D item 1**	**-0.37 vs -0.40**	**No**
4	**The correlation of change on HOOS-PS with EQ-5D item 2 (self-care) is at least 0.10 higher than the correlation of change on HOOS Pain with EQ-5D item 2**	**-0.27 vs -0.24**	**No**
5	**The correlation of change on HOOS-PS with EQ-5D item 3 (usual activities) is at least 0.10 higher than the correlation of change on HOOS Pain with EQ-5D item 3**	**-0.43 vs -0.44**	**No**
6	**The correlation of change on hip specific anchor with HOOS Pain is at least 0.10 higher than the correlation of change on hip specific anchor with EQ-VAS**	**-0.40 vs -0.25**	**Yes**
7	**The correlation of change on hip specific anchor with HOOS-PS is at least 0.10 higher than the correlation of change on hip specific anchor with EQ-VAS**	**-0.39 vs -0.25**	**Yes**
8	**The correlation of change on hip specific anchor with HOOS QoL is at least 0.10 higher than the correlation of change on hip specific anchor with EQ-VAS**	**-0.46 vs -0.25**	**Yes**
9	**The correlation of change on HOOS Pain with EQ-5D item 4 (pain/discomfort) is at least 0.10 higher than the correlation of change on HOOS Pain with EQ-5D item 1 (mobility)**	**-0.52 vs -0.40**	**Yes**
10	**The correlation of change on HOOS Pain with EQ-5D item 4 (pain/discomfort) is at least 0.10 higher than the correlation of change on HOOS Pain with EQ-5D item 2 (self-care)**	**-0.52 vs -0.24**	**Yes**
11	**The correlation of change on HOOS Pain with EQ-5D item 4 (pain/discomfort) is at least 0.10 higher than the correlation of change on HOOS Pain with EQ-5D item 3 (usual activities)**	**-0.52 vs -0.44**	**No**
12	**The correlation of change on HOOS Pain with EQ-5D item 4 (pain/discomfort) is at least 0.10 higher than the correlation of change on HOOS Pain with EQ-5D item 5 (anxiety/depression)**	**-0.52 vs -0.12**	**Yes**
13	**The correlation of change on HOOS Pain with EQ-5D item 4 (pain/discomfort) is at least 0.10 higher than the correlation of change on HOOS Pain with EQ-VAS**	**-0.52 vs -0.34**	**Yes**

n=1,025. Total amount of hypothesis that were rejected: 6/13. Responsiveness considered moderate.
1: Hypothesis formulated after data collection but before data analysis 2: Spearman's correlation coefficients.

this context. In addition, the former definition is limited to postal surveys. It is likely that internet based applications will be used to an even higher degree in the future, and internet based PROs should also be included in a modern definition of feasibility.

Validity and feasibility of a PRO is not absolute but depends on the context in which it is being used. A PRO will therefore not be valid per se, but can be validated in a specific context; for example for THA patients in a Registry setting. PROs validated in similar settings (like RA patients), may contribute to our assessment of the validity of a PRO where no other validations exist. An example of this: the EQ-5D has been validated for both THA patients (187), and for RA patients (120). If no EQ-5D validation on THA patients existed, the results from the validation in RA patients could have been used due to some similarities between THA patients and RA patients. This is of course not ideal (because of the many differences between THA and RA patients), but when lacking information from the patient group in question, similar patient groups could be used, as some information is better than no information.

Response rate

The feasibility criteria in study I included response rate, and our cut-point was a response rate of 80%. All PROs in study I had a response rate over our cut-point (an overall response rate of 83%). In study III, the OHS had an excellent response rate of 87%. Other PRO studies including THA patients, THA and TKA patients combined or patients with revision hip replacement, have found response rates ranging from 62-88% (44;100;188-190). In the Swedish Hip Arthroplasty Registry, response rates of PROs range from 49-92% (7). The low response rate of 49% was achieved with an internet based application, and declining response rates with increasing age was seen. No differences in regard to population density was found (7). The impressive 1 year response rate of 92% with pen-and-paper questionnaires found is comparable to the 96% of patients who answered the 1 year postoperative questionnaire set in our study IV. The difference in response rate in study I (83%) compared to in study IV (96%) might be explained by differences in the methodology; the first study was a register study where a sample of patients from the register were sent both invitations to participate in the study as well as questionnaires. Therefore patients declining study participation were subtracted from the included patients. In study IV (a cohort study) the patients could decline study participation before they were given the PROs, so the patients declining study participation were not subtracted from the included patients. Removing the patients declining to participate from the included patients in Study I would give a response rate of 90%. So, in different settings the term response rate can have similar but different meanings. A sufficiently high response rate is vital to minimize selection bias and to ensure generalizability. A low response rate would increase the risk of selection bias: Rolfson et al. found that using an internet questionnaire alone gave an insufficient response rate and biased results since older patients and those with more severe co-morbidities did not respond (7).

Only few studies have evaluated whether follow-up time affects the response rate in a joint registry context. The New Zealand joint registry chose to send out OHS at six months postoperative, as it was reasoned that the operation and rehabilitation would then still be a recent significant event for the patient and therefore encourage a high response rate. They achieved a 75% response rate at six months, but achieved a five year response rate of 80% (56). In the Cochrane review by Edwards et al., there was found no evidence for an effect on response rate of questionnaires being sent sooner after discharge from hospital, and no evidence for an effect on response rate when a follow-up interval of less than 31 days was used (91). In study I, I saw no difference in response rate depending on follow-up times ranging from 1 to 11 years. This supports the view that follow-up time is not strongly related to response rate, which may be explained by that patient burden and patient-perceived importance have a much higher impact on patients' decision to answer a PRO, than time after surgery. Many factors can contribute to an increased response rate in both postal and electronic surveys, some of them are listed in Table 8 (91;93). I used several of these strategies to achieve our response rate (described in the methodological considerations).

Table 8. Methods to increase response rates of PROs

Methods to increase response rates of PROs
Including an invitation letter
Advance warning that the questionnaire will be coming
Giving a token of appreciation
Enclosing a stamped self-addressed envelope
Follow-ups
Monetary incentives
A teaser on the envelope - e.g. a comment suggesting to participants that they may benefit if they open it
A more interesting questionnaire topic
Unconditional incentives
Shorter questionnaires
Providing a second copy of the questionnaire at follow up
Mentioning an obligation to respond
University sponsorship
Non-monetary incentives
Personalized questionnaires
Use of hand-written addresses
Use of stamped return envelopes as opposed to franked return envelopes
An assurance of confidentiality

Ceiling effect

A ceiling effect of 6-46% was found in study I and III. For all PRO subscales studied, except the SF-12 subscales and the EQ-VAS, the ceiling effect were over the 15 % considered the maximum acceptable (64). This is in accordance with other studies which has showed a similar ceiling effect (101;185;191). SF-12 PCS and SF-12 MCS had lower ceiling effects, as reported by others, which is explained by computation of a norm-based value set (192). The high ceiling effect in the present thesis could be explained by the postoperative administration of the PROs. Considering the median postoperative follow-up period of five years in study III and the good overall clinical outcome from THA (191), it could be argued that the finding is merely a degree of skew, which is to be expected given the timing of measures relative to the intervention (PROs administered postoperatively), and this can explain the skew in sum score distribution. A lower ceiling effect preoperatively compared to postoperatively is self-evident, and previously shown by others (191). Consistent with our findings in study III, Naal et al. found a lower preoperative OHS ceiling effect (107). Considering the good outcome of THA, low floor effects and high ceiling effects can be expected and I contribute the high number of ceiling scores, to the combination of a fairly long postoperative follow-up period in our study I and III, and the good overall clinical outcome from THA; therefore, I believe the proposed and fairly arbitrary criterion of having the best possible score in less than 15% of patients following THA might be too

restrictive in a standard population. The ceiling effect will also be dependent on the preoperative scores. Patients with a very good preoperative score may mistakenly be misclassified as non-responders because their baseline score does not allow achievement of important change due to ceiling effects (193). The responsiveness can be defined as the ability of a PRO to detect change over time in the construct to be measured (3). Ceiling effect may influence the reliability and the responsiveness of a PRO because it is not possible to see if a patient improves or is in the same state for repeated measurements. de Vet et al. argue that if there really is a ceiling effect depends on whether one want to discriminate the patient group any further; after joint replacement a high percentage of patients may have ceiling scores, but they argue that this should not be considered a ceiling effect since one do not want to discriminate these postoperative patients any further (194). This can however be debated. In a registry context a long-time follow-up is important, and in this setting it is therefore preferential to be able to discriminate also postoperative patients based on their PRO outcome. As a consequence of this, a high percentage of ceiling scores should be defined as a ceiling effect, a PRO used by THA patients in a joint registry context should have as small ceiling effect as possible, but a criterion of a postoperative ceiling effect below 15% might be too restrictive.

In study III, the ceiling effect is reported to be 19.9% in both text and table (Paper III, Table III), but in the figure it is 23% (Paper III, Figure 2). The explanation to this apparently discrepancy were unfortunately removed from the paper in the review process; The ceiling effect in the text and table is reported without imputation of missing items (the percent of number of PROs with best possible answer on every item) to make it easier to compare ceiling effects between the different PROs, while in the figure, the ceiling effect is reported with imputation of missing items (the percent of PROs with best possible sum score, after imputation).

In study I, one part of the conclusion was: "We found minor differences between the disease-specific and the generic PROs regarding ceiling and floor effects as well as discarded items". This might be considered to be a controversial statement, since the ceiling effect of the different PROs varied from 6-46%. It is important to note that this statement concerns the difference in ceiling and floor effects between the examined disease-specific PROs compared to the generic PROs (20-37% vs. 6-46% and 0-0.5% vs. 0-0.3%), and in this context the difference can be interpreted as minor. However, all ceiling effects for the different PROs were published, as this information may be useful for decision making about what PRO to include, when a low ceiling effect is of a particular interest.

Missing data

Missing data can be a major challenge in ensuring good PRO data quality (192). Missing data may decrease data quality and have the potential to undermine the validity of the results, if occurring not random. Imputation of missing data can be an option (195). The impact of imputation of missing data in hip replacement patients for OHS and EQ-5D have previously been assessed, and the differences in mean scores between PROs with or without imputation have been found to be very small (109). The handling of missing data by imputation in the included studies is described in the methodological considerations. Imputed data can be problematic to use for assessing the measurement properties of a PRO instrument, as imputed data will artificially reduce variation in overall scores, and this is a known limitation in study III. Study III is a secondary analysis of data from study I, which explains the use of imputed data in this study. In study I the cut-off chosen in regard to an acceptable proportion of items missing was 5%. Others find 3-15% missing items acceptable (184). In study I the proportion of items missing ranged from 1.2-3.4%, and 1.2-5.5% of the PRO subscales had to be discarded due to too many items missing, making it impossible to calculate a sum score. In study III 0.5-4.2% had too many items missing to calculate a sum score. Completion rate (the percentage of PROs with too many items missing, subtracted from the total number of PROs) is in the literature sometimes used to describe missing items. Table 9 show the completion rate of the different PROs in study IV.

Table 9. Additional results: Completion rates (%) for PRO subscales, from Study IV

PROs	Preoperative (n=1,335)	Postoperative (n=1,288)
HOOS Pain	**98.3 (n=1,312)**	**95.7 (n=1,233)**
HOOS PS	**98.3 (n=1,312)**	**96.7 (n=1,245)**
HOOS QoL	**98.8 (n=1,319)**	**97.2 (n=1,252)**
EQ-5D Index	**95.2 (n=1,271)**	**95.3 (n=1,228)**
EQ-VAS	**97.5 (n=1,302)**	**95.6 (n=1,231)**

The proportion of missing items, the percentage of discarded PROs due to missing items and the completion rate is similar in all studies presented, and correspond well with the numbers reported in other PRO studies (100;107;196). There are also studies showing an even smaller percentage of missing items (around 0%) than in the studies presented in this thesis (106;159;160). The number of included patients and how the patients were managed in the follow-up may explain these differences. Rolfson et al. report completion rates of EQ-5D in the Swedish Hip Arthroplasty Registry of 86.1% (preoperative) and 90.2% (postoperative) (7). In study I (58 % females), I found that females left more unanswered items than males. This may partly explain the high amount of missing items in the study of 3,156 RA patients (75-80 % females), where 7 % of patients were missing more than 20 % of items for SF-12 PCS, SF-12 MCS and EQ-5D (192).

For EQ-5D, the percentage of PROs with missing items (missing EQ-5D items in study I per total number of EQ-5D items in study I) is reported for each item and for the EQ-VAS and EQ-5D Index in Table 10.

Table 10. Additional results: Number (%) of PROs with missing items for each item for EQ-5D, from Study I

Question *1*	**n (%)**
1	**28 (1.2)**
2	**34 (1.4)**
3	**30 (1.2)**
4	**19 (0.8)**
5	**36 (1.5)**
Total score *1*	**n (%)**
EQ-VAS	**132 (5.5)**
EQ-5D Index	**76 (3.2)**
n=2,407. 1: No imputing of missing values	

The percentage of PROs with missing items in the EQ-VAS and EQ-5D Index in Table 10 is the same as "Discarded PRO subscales" (Paper I, Table 3), because there is no imputing of missing values for EQ-5D. The different proportion of items missing in study III

(Paper III, Table III) is due to that this proportion is calculated of the total number of PROs (not the total number of PRO subscales). The percentage of PROs with missing items differs from the missing items in study III (Paper III, Table III), as study III only includes a smaller sample (n=898) of the patients included in study I (n=2,407). The percentage of PROs with missing items in study I (Table 10) is comparable to the percentage of PROs with missing items in study IV (Table 11). There seems to be a higher percentage of missing items postoperatively than preoperatively (Table 11).

Table 11. Additional results: Number (%) of PROs with missing items for each item for EQ-5D, from Study IV

Question *1*	**Preoperative (n=1,286)**	**Postoperative (n=1,245)**
	n (%)	**n (%)**
1	**20 (1.6)**	**31 (2.5)**
2	**37 (2.9)**	**31 (2.5)**
3	**19 (1.5)**	**35 (2.8)**
4	**22 (1.7)**	**32 (2.6)**
5	**33 (2.6)**	**44 (3.5)**
Total score *1*	**n (%)**	**n (%)**
EQ-VAS	**31 (2.4)**	**54 (4.3)**
EQ-5D Index	**62 (4.8)**	**59 (4.7)**
1: No imputing of missing values		

For HOOS, the percentage of PROs with missing items (missing HOOS items in study I per total number of HOOS items in study I) is reported for each item and for the HOOS Pain, HOOS-PS and HOOS QoL in Table 12.

Table 12. Additional results: Number (%) of PROs with missing items for each item for HOOS, from Study I

Question *1*	**n (%)**
1	**255 (10.8)**
2	**59 (2.5)**
3	**69 (2.9)**
4	**74 (3.1)**
5	**60 (2.5)**
6	**51 (2.2)**
7	**62 (2.6)**
8	**75 (3.2)**
9	**61 (2.6)**
10	**58 (2.5)**
11	**74 (3.1)**
12	**67 (2.8)**
13	**74 (3.1)**
14	**189 (8.0)**
15	**79 (3.3)**
16	**64 (2.7)**
17	**49 (2.1)**
18	**50 (2.1)**
19	**47 (2.0)**
Total score *2*	**n (%)**
HOOS Pain	**72 (3.0)**
HOOS-PS	**64 (2.7)**
HOOS QoL	**44 (1.9)**
n=2,365. 1: No imputing of missing values 2: Imputing of missing values	

The percentage of PROs with missing items in the HOOS Pain, HOOS-PS and HOOS QoL in Table 12 is the same as "Discarded PRO subscales" in study I (Paper I, Table 3), because these sum scores are calculated after imputation. The different proportion of items missing in study III (Paper III, Table III), is due to that this proportion is calculated of the total number of PROs (not the total number of PRO subscales). The percentage of PROs with missing items is comparable but different from the missing items in study III (Paper III, Table III), as study III only includes a smaller sample (n=187) of the patients included in study I (n=2,365). The percentage of PROs with missing items in study I (Table 12) is comparable to the percentage of PROs with missing items in study IV (Table 13), but there seems to be a higher percentage of missing items postoperatively than preoperatively (Table 13).

Table 13. Additional results: Number (%) of PROs with missing items for each item for HOOS, from Study IV

Question *1*	**Preoperative (n=1,286)**	**Postoperative (n=1,245)**
	n (%)	**n (%)**
1	**56 (4.4)**	**140 (11.2)**
2	**25 (1.9)**	**49 (3.9)**
3	**23 (1.8)**	**50 (4.0)**
4	**25 (1.9)**	**51 (4.1)**
5	**22 (1.7)**	**49 (3.9)**
6	**15 (1.2)**	**49 (3.9)**
7	**19 (1.5)**	**46 (3.7)**
8	**22 (1.7)**	**59 (4.7)**
9	**19 (1.5)**	**50 (4.0)**
10	**15 (1.2)**	**45 (3.6)**
11	**22 (1.7)**	**38 (3.1)**
12	**23 (1.8)**	**42 (3.4)**
13	**26 (2.0)**	**46 (3.7)**
14	**62 (4.8)**	**103 (8.3)**
15	**23 (1.8)**	**48 (3.9)**
16	**13 (1.0)**	**46 (3.7)**
17	**17 (1.3)**	**37 (3.0)**
18	**17 (1.3)**	**37 (3.0)**
19	**14 (1.1)**	**39 (3.1)**
Total score *2*	**n (%)**	**n (%)**
HOOS Pain	**22 (1.7)**	**51 (4.1)**
HOOS-PS	**20 (1.6)**	**42 (3.4)**
HOOS QoL	**13 (1.0)**	**35 (2.8)**
1: No imputing of missing values 2: Imputing of missing values		

For HOOS, two items seem to have higher percentages of missing items then the rest of the items in study I and in study IV (both preoperatively and postoperatively); the item "How often is your hip painful?" (Answer options; Never, Monthly, Weekly, Daily, Always), and the item "The following questions concern your level of function in performing usual daily activities and higher level activities. For each of the following activities, please indicate the degree of difficult you have experienced in the last week due to your hip problem; Running" (Answer options; None, Mild, Moderate, Severe, Extreme). There may be many potential explanations to why these items often are missed, including the layout of the questionnaire and items, the number and content of answer options, and the patient-perceived relevance of the items. Both items have good face validity, seem relevant and seem to have sufficient answer options (5 answer options to all HOOS items). I have shown that patients do not have much pain, at least at the postoperative follow-up, but this is included in the answer

options (Never). Pain localization may be a problem, especially in older cohorts with high levels of co morbidity. The pain item is the first item in the HOOS questionnaire used, and I cannot exclude that the placement of the item on the questionnaire could have contributed to it being overlooked by the patients. The item concerning running may have a lower patient-perceived relevance in my population. HOOS was constructed to get improved validity for those with less severe disease or higher demands of physical function, but since the median age in study I and study IV were 68 and 73 respectively, and because of co morbidity, running may not be that relevant for this patient group.

For OHS, the percentage of PROs with missing items (missing OHS items in study I per total number of OHS items in study I) is reported for each item and for the OHS sum score in Table 14.

Table 14. Additional results: Number (%) of PROs with missing items for each item for OHS, from Study I

Question *1*	**n (%)**
1	**25 (1.0)**
2	**14 (0.6)**
3	**12 (0.5)**
4	**12 (0.5)**
5	**28 (1.2)**
6	**21 (0.9)**
7	**48 (2.0)**
8	**34 (1.4)**
9	**35 (1.4)**
10	**37 (1.5)**
11	**43 (1.8)**
12	**28 (1.2)**
Total score *2*	**n (%)**
OHS sum score	**30 (1.2)**
n=2,419. 1: No imputing of missing values 2: Imputing of missing values	

The percentage of PROs with missing items in the OHS sum score in Table 14 is the same as the proportion of items missing in study I (Paper I, Table 3), the same as in "Discarded PRO subscales" (no subscale division of OHS) in study I (Paper I, Table 3), and the same as in study III (too many items missing to calculate a sum score) (Paper III, Table III), because the amount of discarded PROs equals the result after imputation.

Table 15. Additional results: Number (%) of PROs with missing items for each item for SF-12, from Study I

Question *1*	**n (%)**
1	**47 (2.0)**
2	**30 (1.3)**
3	**54 (2.3)**
4	**41 (1.7)**
5	**58 (2.4)**
6	**68 (2.9)**
7	**88 (3.7)**
8	**54 (2.3)**
9	**51 (2.1)**
10	**63 (2.7)**
11	**65 (2.7)**
12	**43 (1.8)**
Total score *2*	**n (%)**
PCS & MCS	**55 (2.3)**
n=2,377. 1: No imputing of missing values 2: Imputing of missing values	

For SF-12, the percentage of PROs with missing items in the PCS and MCS in Table 15 is the same as the proportion of items missing in study I (Paper I, Table 3), and the same as in "Discarded PRO subscales" in study I (Paper I, Table 3). It differs from the missing items in study III (Paper III, Table III), as study III only includes a smaller sample (n=907) of the patients included in study I (n=2,377).

One part of the conclusion in study I might be considered to be a controversial: "We found minor differences between the disease-specific and the generic PROs regarding ceiling and floor effects as well as discarded items". This statement should not be interpreted to mean that there are no differences in the amount of discarded PRO subscales due to missing items between the PROs included in study I, but only that the difference in discarded PRO subscales between the examined disease-specific PROs compared to the generic PROs (1.2-3% vs. 2.3-5.5%) can be interpreted as being minor. The detailed results concerning missing items (Table 10-15) may be useful information in further studies where the number of missing items is of importance.

Manual validation

In this thesis manual validation was defined as an active code validation by a human operator. Manual validation of the scanned PROs was conducted when the AFP system could not convert an answer due to poor or ambiguous questionnaire completion. Therefore, a higher percentage of PRO items needing manual validation may indicate a less patient-friendly PRO format. In need of manual validation the scanner cannot scan further until a human operator manually validates the correct code for the questionnaire answer in question. In Study I, the EQ-5D required manual validation about 3 times as often as the other PROs (Paper I, Table 3), and EQ-VAS had over 4 times the percentage of manual validations per PRO subscales as EQ-5D Index (Table 16), suggesting that the EQ-VAS could be less optimal for use in a mailed survey in a registry population.

Table 16. Additional results: Proportions (Mean, 95% CI) of manual validations per PRO subscales, from Study I

Specific PROs		
HOOS (n=2,365)	**Pain**	**4.3 (3.5-5.1)**
	PS	**2.6 (2.0-3.3)**
	QoL	**2.5 (1.9-3.1)**
OHS (n=2,419)		**7.2 (6.2-8.2)**
Generic PROs		
SF-12 (n=2,377)	**PCS**	**7.7 (6.6-8.8)**
	MCS	**7.7 (6.6-8.8)**
EQ-5D (n=2,407)	**EQ-5D Index**	**4.1 (3.3-4.9)**
	EQ-VAS	**18.8 (17.3-20.4)**

But this finding could also be explained by the use of different AFP technology; in study II, I also found a very high percentage of manually validated items of EQ-VAS compared to the other PROs (about 10 times as often), and I will argue that this is mainly because of ICR. It is clearly more difficult for the AFP system to identify a hand-printed character (number) correctly, than to identify if a check box is marked, also suggested by the higher number of errors per 10,000 data fields in ICR compared to OMR (Paper II, Table 3). This finding is in agreement with older studies, and Jørgensen et al. found it advisable to avoid numeric fields if it could not be assured that respondents would adhere to the recommendations on how to write characters to enhance

recognition (197). Further improvements in ICR technology could possibly decrease the error level to the level of OMR, but this has to be examined in future studies.

The differences in proportions of items validated or the proportion of PROs requiring manual validation also have economic implications. Especially in large studies or in a registry context, where the number of patients (and the number of PROs) are high, the importance of a minimal amount of manual validation is clear. The type of AFP technology required for the different PROs also seem to affect amount of manual validation needed. The data presented in Study I (Paper I, Table 3) can be used to choose PROs requiring the least manual validation possible.

Cost of AFP

Cost will often be an issue when considering implementations of new technologies. Health-economic aspects with health-economic evaluations, cost-per-patient calculations, cost-utility calculations and willingness-to-pay calculations are now an integrated part of the public health system in general and the Hip Arthroplasty Registries (7). AFP technologies have traditionally been expensive and in earlier reports between 54,000 and 99,000 forms are found to be needed to be processed to recover the initial investment (197). The cost of equipment for AFP data capture has decreased considerably in the last decades. The cost of double manual data entry can be very high if the number of questionnaire forms are high, as the time needed for double manual data entry, even with the use of modern data entry software, is substantiel. I found that the mean time for double manual data entry was over two minutes per queestionnaire (Table 17), which would give over 366 effective working hours if the data from Study I alone were to be manually entered.

Table 17. Additional results: Number of seconds (Mean, 95% CI) used for manual double-key data entry *1*, from Study II

PROs	Per PRO	Per Item
HOOS (n=99)	**117 (101-132)**	**6.1 (5.3-6.9)**
OHS (n=100)	**71 (61-82)**	**5.9 (5.1-6.8)**
SF-12 (n=100)	**169 (105-233)**	**14 (8.8-19)**
EQ-5D (n=99)	**195 (92-299)**	**33 (15-50)**
Mean for the entire study (n=398)	**138 (95-182)**	**15 (8.2-21)**

1: Manual double-key data entry (with limiting definitions for entry of out of range values and program control of data entered) using EpiData Data Entry software (EpiData Association, http://www.epidata.dk)

Table 18. Additional results: The cost of AFP vs. manual double-key data entry in Study I

The cost of AFP *1*	**143,772,- DKR**
The cost of manual double-key data entry *2*	**110,032,- DKR**

1: The cost of AFP including: set up and adjusting PROs for AFP, control of status of patients (living or dead) before sending out the PROs, communication with the printing company, printing expenses, sorting of questionnaires and patient information, stapling of questionnaires, mail merging, enveloping, recieving and opening envelopes from the patients, sorting the PROs, removing the staples, scanning the PROs, manual validation of the PROs, sending the data in electronic format, manual checking of out of range values, control of status of patients (living or dead) before sending out the reminders, and managing first and second reminder letters 2: The calculated cost using 4,784 patients, 2 PROs each (as in study I), 138 seconds per PRO entry, and a hourly rate of 300,- DKR

Reports have shown that AFP can reduce processing time to about one half to one third of that of manual data entry and that wage expenses can be reduced to about one third to one quarter (197). Already in 2001, Weller et al. concluded that their AFP system were cost-effective (198). The cost of AFP has not been directly compared to double manual data entry in our studies. The cost of AFP in our study I is listed in Table 18. It may seem as the AFP in study I was not cost reducing, but the cost of AFP included other necessary expenses, and I therefore believe that the AFP was in reality cost reducing. In the comparison in Table 18, a manual hourly rate of 300,- DKR is used, as this is the hourly rate for AFP paid in these studies. The cost of manual data entry will of course depend on the hourly rate of the personnel doing the manual data entry. A higher percentage of PRO items needing manual validation are more costly due to the manual labor required. A British report have shown the overall average cost per matching preoperative and follow-up questionnaire to be £5.81 per patient (109) (approximately 50,- DKR), which is comparable to the cost in this study (Table 18), considering that only postoperative questionnaires were used in study I. Even though the cost of equipment for AFP data capture has decreased considerably in recent decades, substantial time and computer expertise is still required for implementation. Further studies should assess the cost of modern AFP systems in direct comparison to double manual data entry.

Factor analyses

For OHS only 1 factor had an eigenvalue of >1 (11.7), and this factor could explain >99% of the variance in the data set. Large difference in eigenvalues between factor 1 to 2 and small difference in eigenvalues between 2 and 3, and the scree plot supported unidimensionality (Table 19) (199). Oblique rotation did not add clarity to the result already found. OHS was developed to assess function and pain in patients undergoing THA (101). OHS was found to be unidimensional, and measure hip problems in THA patients.

For HOOS Pain only 1 factor had an eigenvalue of >1 (9.9), and this factor could explain >99% of the variance in the data set. Large difference in eigenvalues between factor 1 to 2 and small difference in eigenvalues between 2 and 3, and the scree plot supported unidimensionality (Table 19) (199). Oblique rotation did not add clarity to the result already found. HOOS Pain was found to be unidimensional, and measure hip pain in THA patients.

For HOOS-PS only 1 factor had an eigenvalue of >1 (4.9), and this factor could explain >99% of the variance in the data set. Large difference in eigenvalues between factor 1 to 2 and small difference in eigenvalues between 2 and 3, and the scree plot supported the unidimensionality previously reported (Table 19) (199;200). Oblique rotation did not add clarity to the result already found. HOOS-PS was found to be unidimensional, and measure hip-related physical function in THA patients.

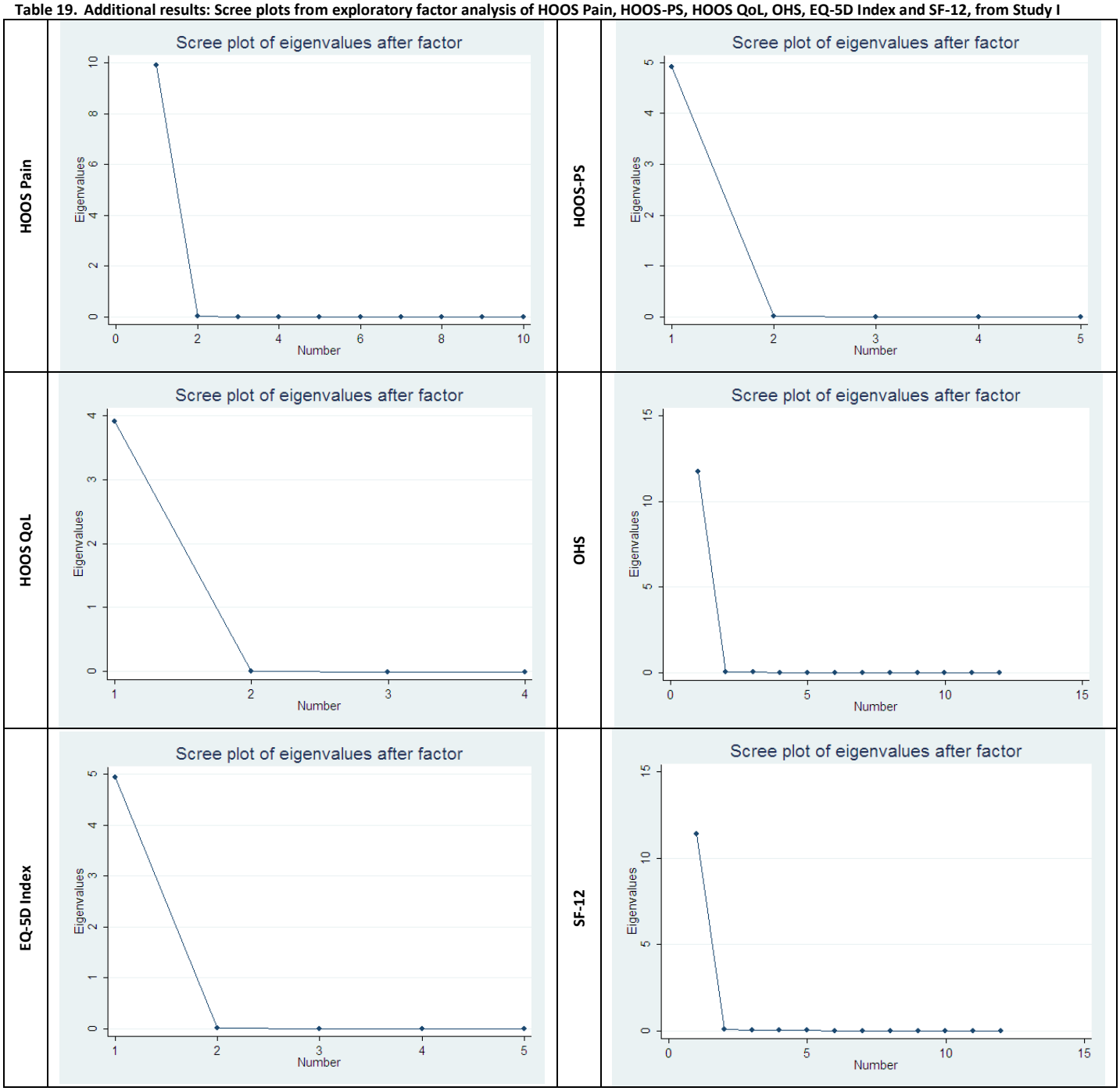
For HOOS QoL only 1 factor had an eigenvalue of >1 (3.9), and this factor could explain >99% of the variance in the data set. Large difference in eigenvalues between factor 1 to 2 and small difference in eigenvalues between 2 and 3, and the scree plot supported unidimensionality (Table 19) (199). Oblique rotation did not add clarity to the result already found. Nilsdotter et al. do not describe the different factors of HOOS in detail, but only state that “all items loaded on a major factor for each subscale” (201).

HOOS QoL was found to be unidimensional, and measure hip-related quality of life in THA patients.

For SF-12 only 1 factor had an eigenvalue of >1 (11.4), and this factor could explain >98% of the variance in the data set. Large difference in eigenvalues between factor 1 to 2 and small difference in eigenvalues between 2 and 3, and the scree plot supported unidimensionality (Table 19) (199). Oblique rotation did not add clarity to the result already found. SF-12 was developed to assess general health (GH), physical functioning (PF), role physical (RP), role emotional (RE), bodily pain (BP), mental health (MH), vitality (VT) and social functioning (SF) (173). The eight hypothesized factors could not be identified. A poor dimensional reproducibility for SF-36 has previously been reported (202). SF-12 was found to be unidimensional, and in this context seems to measure hip-related problems in general health of THA patients.

For EQ-5D only 1 factor had an eigenvalue of >1 (4.9), and this factor could explain >99% of the variance in the data set. Large difference in eigenvalues between factor 1 to 2 and small difference in eigenvalues between 2 and 3, and the scree plot supported unidimensionality (Table 19) (199). Oblique rotation did not add clarity to the result already found. EQ-5D was developed to assess mobility, self-care, usual activities, pain/discomfort and anxiety/depression (172). EQ-5D was found to be unidimensional, and in this context seems to measure hip-related problems in general health of THA patients.

Table 19. Additional results: Scree plots from exploratory factor analysis of HOOS Pain, HOOS-PS, HOOS QoL, OHS, EQ-5D Index and SF-12, from Study I



Differential item functioning

Differential item functioning (DIF) exists when item responses by members of different groups are statistically different when controlling for trait and may indicate item bias. DIF is classified as either uniform (if the effect is constant) or non-uniform (if the effect varies conditional on the trait level) (203;204). Some items (time since operation; item 2, 3, 4, 7. age group; item 1, 3, 5, 6, 7, 10, 12. sex; 2, 4, 5, 7, 12) were initially flagged for uniform DIF from the criterion of a significant LR-test (group as explanatory variable). Large dataset are known to produce significant results even where no clinical relevant differences exist (205). It is common to use a minimum 10% change in effect size (beta) as a criteria for clinically relevant DIF. Using this cut-off, there was no uniform DIF. Furthermore, there was no non-uniform DIF for time since operation or sex, however, for age group a clinically relevant non-uniform DIF was found for item 1 (chi square-p=0.0003) (Table 20).

Table 20. Additional results: DIF of OHS items, from Study III

Subgroups	Item number for items initially flagged for uniform DIF *1*	Item number for items with clinically relevant uniform DIF *2*	Item number for items with clinically relevant non-uniform DIF
Time since operation	**2, 3, 4, 7**	**none**	**none**
Age group	**1, 3, 5, 6, 7, 10, 12**	**none**	**1 (p = 0.0003)**
Sex	**2, 4, 5, 7, 12**	**none**	**none**

1: Significant LR test 2: 10% change in effect size

For the WOMAC pain subscale, age-related DIF have been reported in hip OA patients (206). There may be several reasons why age group affect the pain reported in the OHS item: "During the past 4 weeks… How would you describe the pain you usually had from your hip?" (Answer options; None, Very mild, Mild, Moderate, Severe). There is no clear consensus in the literature on how age affect pain after THA; Older age have been associated with moderate-severe pain 2-5 years postoperative revision THA (207). Others report no age related differences in joint pain 6 month postoperative THA or TKA (208). In a review of 64 THA and TKA studies, Santaguida et al. concluded that age do not influence the outcome of pain (209). I have found that the PASS concerning pain is different for different age groups (Study IV, Table 4), implying that the amount of pain found acceptable changes with age, which may partly explain the findings. The time of measure (less than 1 year follow-up postoperatively) and the type of pain measure itself may also affect the results.

PRO completion in same state

Patients who have not changed and have the same health state in repeated measurements are defined as being in same state. For example: preoperative and postoperative THA patients are not in same state (due to the THA), but patients who have received a THA several years previously and then completes a PRO twice (with only two weeks in between) may very well be in a same state. If a patient in this situation scores different on the PROs, this may have two explanations: either the patient is not in a same state, or there is a problem with the PRO (poor test-retest reliability). If the patient scores the same on the PRO twice, there are also two possible explanations: either the patient is in a same state, or the PRO cannot capture the change in the patient (for instance due to floor- or ceiling effect or poor responsiveness). If a given PRO has not been completed when patient was in the same state, this will lessen the strength of comparisons such as correlations between related constructs. This may be a limitation in study III and the OHS calculations, since I do not know whether the PROs have been completed in same state. Ideally the PROs should have been completed simultaneously when assessing internal consistency and test-retest reliability. One the other hand, if the patients complete a PRO twice in close succession they may remember their last answers and answer the same, instead of answering in accordance to how they feel. Test-retest intervals of 24 hours to four weeks have been used (93;99;210) and a medium test-retest interval of two weeks was therefore chosen. To minimize day to day variations, each item in OHS is started with 'During the past 4 weeks…', the same time period is used in SF-12, and in HOOS a one week time period is used. In study I, both questionnaires were sent to the patients in the same envelope, and returned from the patients together in another envelope. I therefore assume that most questionnaires were completed the same day. Since the patients cannot complete two questionnaires simultaneously, an assumption of a reasonably same state is necessary. In study I and study II, I assume that patients were in the same state, regarding their hip, since the patients are postoperative, and mostly beyond one year follow-up, in study III; 0.9-10.5 years postoperative (4.9 years median, 5.0 years mean).

MCII and PASS estimation

Few estimates of MCII and PASS for different PROs used in orthopedic surgery have been published (124;131). The present thesis presents estimated MCII and PASS at 1 year following THA for the HOOS Pain, the HOOS PS, the HOOS QoL, the EQ-5D Index and the EQ-VAS (Paper IV, Table 1-3). MCII and PASS are of importance because they represent cut-point values for the minimal clinically important improvements in PRO change scores, and cut-point values for the postoperative PRO score found acceptable by the patients, and focus on the patient perspective of outcome. Study IV showed that it is possible to determine cut-off points for the change considered representing the MCII and for the postoperative score considered represent the PASS following THA. After THA, an improvement of approximately 38-55% from mean baseline PRO score and an absolute follow-up scores of 57-91% of the maximum score corresponded to MCII and PASS, respectively. Earlier MCII estimates for EQ-5D varies considerably (0-0.69; in patients with RA, psoriatic arthritis and ankylosing spondylitis after 3 months of treatment with disease-modifying antirheumatic drugs) implying that MCII is dependent on patient group (138). Further strengthening this assumption, Walters et al. found that the mean MCII for the EQ-5D Index was on average 0.074 (range -0.011-0.140) in 8 longitudinal studies with 11 patient groups (no hip OA or THA patients included) (150), pointing towards that there are no universal cut-points for a single PRO and that the estimates will vary by population and context (132). Tubach et al. estimated MCII to be 15 of 100 for absolute improvement, for 4 different generic PROs in chronic rheumatic diseases (RA, ankylosing spondylitis, hand osteoarthritis, hip and/or knee osteoarthritis, and chronic back pain) in a multinational cohort study of 1,532 patients (211). This low MCII estimate can be explained by the different patient population (chronic rheumatic diseases vs. THA patients), the different intervention (nonsteroidal anti-inflammatory drugs vs. THA) and the different follow-up time (4 weeks vs. 1 year). Our finding of an MCII of 0.31 for the EQ-5D Index corresponds well with previous findings of 0.32 (anchor based methods, identical

anchors and estimation approach) and 0.42 (distribution based methods) for THA patients 6 months after surgery (109).

Also PASS estimates for EQ-5D Index exist (0.70; RA patients after 3 months of treatment with disease-modifying antirheumatic drugs) (138), and PASS has been reported to be 40 of 100 for absolute improvement in patients with chronic rheumatic diseases) (211). The low PASS may be explained by the different intervention and inclusion of a different patient population (patients with a chronic disease), compared to Study IV. Regarding PASS, patients' expectations and threshold for an acceptable symptom state may be higher in THA patients (due to the intervention itself) and due to the nature of their disease; chronic rheumatic patients (and patients with severe symptoms) may have a lower threshold for an acceptable symptom state (181;211). PASS estimates is known to vary depending on estimation approach (212), the methodology for identification of PASS have been found to influence the identified cut-points and the ROC approach generally provide lower estimates than the cut-points identified with the 75th percentile approach (139).

Davies et al. have stressed the importance of estimating PASS scores in tertiles (lowest-, middle-, and highest subscale scores) of the preoperative PRO scores, as the baseline score may not allow achievement of important change for patients with the lowest preoperative score (193). MCII has been shown to vary more across tertiles of baseline scores than PASS (181). Browne et al. found that in general, there was little association between baseline severity and MCII values, but recommend to test for this association when generating anchor-based MCIIs from change scores (213). Other cut-points might have been found using other estimation methods; our results should be interpreted with caution, and considered contributing to the emerging knowledge on interpretation of PRO scores in orthopedics. The results from the different methods found in study IV points toward the validity of the cut-points found. Subgroups of different sex, diagnoses and age may have different MCII and PASS (214). In study IV, I found that males had better PASS estimates than females, idiopathic OA patients better PASS estimates for HOOS QoL and EQ-5D Index than other patients and patients over 70 years had lower PASS estimates than younger patients for HOOS Pain, HOOS-PS and EQ-VAS (Paper IV, Table 4). MCII for subgroups were not estimated due to small subgroups, but mean PRO change scores for the different subgroups were calculated (Table 21). MCII subgroup estimations on the same PROs and patient group are therefore warranted in further studies. In large randomized clinical trials result can be statistically significant without being clinically relevant and estimated PASS and MCII can help in the interpretation of data in these kinds of studies, as well as in registry studies. Study IV is to our knowledge the first MCII and PASS estimation study for the HOOS and EQ-5D with THA patients.

Anchors and anchor-PRO correlations

The view presented in this thesis is that anchor-based methods is the only way to estimate MCII and PASS based on the patients perspective, and therefore anchor-based methods should be used when the focus is the patients perspective. Two different approaches for estimating MCII and PASS have been described: Anchor-based methods and distribution based methods. Global transition questions and clinical anchors are different types of anchor-based methods. Standard error of measurement and effect size are examples of distribution based methods. Anchor-based methods (and distribution based methods) both have advantages and disadvantages: Anchor-bases data are often easy to obtain and may provide external basis for interpretation, but may be influenced by wording of PROs and anchors, and an adequate anchor-PRO correlation is required, as summarized by Crosby et al., who advocate the use of both anchor-based and distribution based methods (161). Revicki et al. found that the anchor-based methods should be preferred, with the distribution based approaches providing supportive evidence (133). This is in agreement with other recommendations: King recommends that multiple methods should be used to determine MCII with global transition questions and clinical anchors providing primary evidence, using SEM and ES as supportive evidence (132). Guyatt et al. conclude in a similar fashion; distribution based methods will not suffice on their own, but will be useful to the extent that they bear a consistent relationship with anchor-based methods (215). As previously mentioned, the wording of the anchor questions is also important: Barber et al. found that the wording of the anchor affected the interpretation of change in the PRO score, and that different anchors lead to different estimation results (216).

Multiple anchor-based approaches were used to estimate MCII and PASS in study IV. In addition to the multiple estimation approaches, several anchors (type of anchor questions) were included. In study IV, both hip-specific anchors and general health anchors were used. I chose to estimate MCII and PASS for the hip-specific HOOS based on both hip-specific- and general health anchors. By using hip-specific- and general health anchors one examines different concepts. The concept examined by using a hip-specific anchor is hip related pain, physical function and quality of life after THA, while the concept examined by using a general health anchor is the impact of THA on general health. When reporting hip improvement, patients may value a smaller change in PRO score important than the change in PRO score

Table 21. Additional results: PRO change scores (mean, 95% CI) for different subgroups of sex, diagnoses and age, from Study IV

PRO	Males	Females	p-value *1*	Idiopatic OA	Other diagnoses	p-value *1*	<50 years	50-70 years	>70 years	p-value *2*
Δ HOOS Pain	**45 (43-46)**	**44 (42-46)**	**0.682**	**45 (44-46)**	**43 (38-49)**	**0.642**	**45 (40-50)**	**46 (45-48)**	**42 (40-44)**	**0.003**
Δ HOOS PS	**44 (42-45)**	**42 (40-44)**	**0.359**	**43 (42-44)**	**42 (36-47)**	**0.645**	**43 (38-49)**	**45 (43-47)**	**40 (38-42)**	**0.001**
Δ HOOS QoL	**49 (47-51)**	**48 (46-50)**	**0.319**	**49 (48-51)**	**43 (36-49)**	**0.055**	**47 (41-54)**	**50 (48-52)**	**47 (44-49)**	**0.087**
Δ EQ-5D Index	**0.27 (0.25-0.28)**	**0.28 (0.26-0.30)**	**0.408**	**0.27 (0.26-0.29)**	**0.27 (0.22-0.33)**	**0.974**	**0.24 (0.19-0.29)**	**0.30 (0.28-0.31)**	**0.24 (0.22-0.26)**	**0.001**
Δ EQ-VAS	**18 (16-20)**	**18 (16-20)**	**0.741**	**18 (17-20)**	**18 (12-23)**	**0.875**	**25 (19-32)**	**20 (18-22)**	**14 (12-17)**	**<0.001**

1: Welch's t-test 2: W-test

required to make an impact on the minimal important improvement in the patients' general health. I estimated MCII and PASS for the general health focused EQ-5D based on general health anchors. It can be argued that there is no scientific rationale to compare HOOS data to the general health anchor question, since a hip-specific questionnaire cannot be used to determine general health. On the contrary, hip surgery can affect both hip-specific and general quality of life so there is a rationale to compare both HOOS and EQ-5D with the hip-specific anchor question.

Anchor-based differences can be determined either longitudinally (change in score of one group over time) or cross-sectionally (differences between clinically-defined groups at one time point), and the MCII results presented (Study IV, Table 2) were determined longitudinally. A global transition question is an item which requires patients to remember a prior health state and compare it to how they are currently feeling. In longitudinal studies it is the most commonly used anchor-based method for determining the MCII (132;158). A fifteen-point scale, a seven-point scale or a five-point scale can be used (158;217). 5-9 possible answers have been reported to be considered optimal (93), and all included anchor questions in study IV had a five-point scale. The transition anchor had the five-point scale often used: 'Much better', 'A little better', 'About the same', 'A little worse' and 'Much worse' (132). The patients answers to the five-point scale anchor questions fitted very well with all postoperative PRO scores, with almost no overlap in CI (Table 22 and Table 23).

In the MCII estimation both a retrospective transition anchor and an absolute change (postoperative value - baseline value) anchor (154;166) were used, to reduce the recall bias known to be a problem for retrospective anchors (132;154). Retrospective estimation of health status may be influenced by mood, memory, and attitude (218), and Barber et al. found that the MCII estimates were different when using retrospective- and absolute change anchors (216). In study IV the MCII cut-point estimates were substantially lower for the retrospective anchor, but since there were only one general health anchor and hip-specific anchor for the MCII and PASS estimations, no direct comparison of the retrospective- and absolute change anchor could be done.

The use of an absolute change anchor can be criticized, but several factors support the use of this anchor. EQ-VAS is also a general health item, and this anchor-item correlation is adequate (Paper IV, Supplementary data, Table 8). The mean change score of both EQ-5D Index and EQ-VAS increases in line with better anchor question answer options, with almost no overlap in CI (Paper IV, Supplementary data, Table 7). There also seem to be a connection between the different answer categories and the percentage of patients that decline in usage of nonprescription pain medication, the percentage of patients improving their activity level and the percentage of patients who report improvement regarding strenuousness in daily activities (Table 24). The correlation was >0.5 between the two general health anchors (Spearman's rho = -0.52), and between the two hip-specific anchors (Spearman's rho = 0.54), which strengthen the validity of the anchors used (217). Item 1 of the SF-12 used in study I, is identical with the general health anchor used for PASS estimation in study IV, and the Spearman's correlation coefficients for this item and the HOOS subscales is almost identical (Table 25).

The definition used for MCII is similar to the definition of the minimal clinically important difference, except that MCII only addresses the direction of improvement and not worsening (129). In study IV, patients responding "A little better" were considered as reporting a minimal clinically important improvement (132),

Table 22. Additional results: Postoperative PRO scores (mean, 95% CI) and distribution of different answer categories for the hip specific anchor question; "How would you describe the results of your operation?", from Study IV

Anchor	n (%)	HOOS Pain	HOOS-PS	HOOS QoL	EQ-5D Index	EQ-VAS
Excellent	650 (53)	96 (95-97)	93 (92-94)	91 (90-92)	0.95 (0.94-0.96)	86 (85-87)
Very good	327 (27)	88 (87-90)	85 (83-86)	77 (74-79)	0.85 (0.84-0.87)	79 (77-81)
Good	142 (12)	78 (75-80)	73 (70-76)	66 (61-68)	0.77 (0.75-0.80)	72 (69-74)
Fair	56 (5)	60 (55-65)	56 (51-61)	43 (38-47)	0.66 (0.62-0.70)	63 (58-67)
Poor	42 (3)	55 (47-63)	46 (38-54)	30 (23-37)	0.56 (0.50-0.63)	48 (42-53)

Table 23. Additional results: Postoperative PRO scores (mean, 95% CI) and distribution of different answer categories for the general health anchor question; "In general, would you say your health is...", from Study IV

Anchor	n (%)	EQ-5D Index	EQ-VAS
Excellent	144 (12)	0.98 (0.97-0.99)	95 (94-97)
Very good	429 (35)	0.93 (0.92-0.94)	87 (86-89)
Good	426 (35)	0.88 (0.87-0.89)	79 (78-81)
Fair	180 (15)	0.73 (0.71-0.76)	62 (59-64)
Poor	42 (3)	0.49 (0.41-0.58)	41 (35-47)

Table 24. Additional results: Change in Pain medication usage, Activity level and Strenuousness in daily activities and the distribution of preoperative answers compared to postoperative answers, for the general health anchor question; "In general, would you say your health is...", from Study IV

Anchor	Pain medication usage *1* n (%)			Activity level *2* n (%)			Strenuousness in daily activities *3* n (%)		
	Worse	No change	Better	Worse	No change	Better	Worse	No change	Better
>1 step better	2 (2)	67 (52)	59 (46)	12 (10)	73 (58)	41 (33)	19 (15)	14 (11)	94 (74)
1 step better	18 (4)	219 (54)	169 (42)	33 (8)	278 (70)	89 (22)	94 (23)	82 (20)	228 (56)
No change	25 (5)	274 (56)	187 (38)	49 (10)	344 (71)	89 (18)	119 (24)	113 (23)	256 (52)
1 step worse	7 (5)	100 (68)	40 (27)	36 (25)	94 (64)	16 (11)	58 (40)	27 (19)	60 (41)
>1 step worse	0 (0)	13 (76)	4 (23)	3 (19)	12 (75)	1 (6)	6 (40)	5 (33)	4 (27)

1: Preoperative- and postoperative patient reported nonprescription pain medication usage 2: Preoperative- and postoperative patient reported activity level 3: Preoperative- and postoperative patient reported strenuousness in daily activities

Table 25. Additional results: Spearman's correlation coefficients, HOOS and General health, from Study I and Study IV *1*

PRO	Study I	Study IV
	General health item	General health item
	n=1,016	n=1,179
HOOS Pain	-0.46	-0.45
HOOS-PS	-0.46	-0.46
HOOS QoL	-0.43	-0.41

1: Item 1 of the SF-12 used in study I is identical with the general health anchor used for PASS estimation in study IV: "In general, would you say your health is"

but where the Area Under the Curve (AUC) estimation of the ROC curves (219) were below the proposed minimum of 0.523 (138), patients responding "A little better" or "Much better" were combined, and the results were similar (Study IV, Table 2). If minimal clinically important difference is the focus, both patients getting a little better or a little worse constitute the minimal change subgroup (133), but the change related to an improvement is not necessarily the same change as that for a decline (only negative), so the reported MCII in this thesis is based solely on improvement. A reasonable number got "A little better", and few patients got worse (Study IV, Supplementary data, Table 6).

In study IV, the hip-specific anchors were regarded as most important for the hip-specific PRO (133;134), since there should be a theoretical basis for the relationship between the anchor and the PRO, PRO subscale or relevant domain, and an empirical correlation between the anchor and the PROs included of at least 0.30 (132). The correlation between both the EQ-5D subscales and the hip-specific MCII anchor in study IV were less than 0.30, illustrating the suboptimal correlation between the general health focused EQ-5D and the hip-specific MCII anchor. Also the correlation between the EQ-5D Index and the general health MCII anchor were less than 0.30. This limitation of the EQ-5D Index MCII estimation might be explained by problems with the time-trade-off procedure of the EQ-5D Index (171). The correlation between the HOOS subscales and the general health MCII anchor ranged from 0.25-0.28, and the low correlation illustrate that a hip-specific questionnaire should not be used to determine general health. For all other MCII estimations in study IV, the anchor-PRO correlations were moderate (>0.30), but below 0.50. The anchor-PRO correlations of the PASS estimations were large (>0.50) for all PROs, except for the EQ-VAS´s moderate correlation with the hip-specific anchor (-0.48), which strengthen the presented PASS estimates. In study IV, I have shown that patients who reported most improvement in general health, had the best change score for EQ-VAS and patients who reported least improvement in general health had the worst change score for EQ-VAS, illustrating the acceptable correlation (Paper IV, Supplementary data, Table 7 and Figure 4). Revicki et al. stress the importance of determining this strength of the association, since an anchor that has a very low (or even moderate) correlation may provide misleading information in defining what is important to patients (155), and often will yield estimates that are too small (158). A single-anchor approach will require a higher degree of correlation than a multiple anchor approach (215). The anchor-PRO correlation should be reported in studies (154), but due to the lack of anchor-PRO correlation reported in the literature for the PROs and patient group included, comparison to other studies is difficult.

Anchor–based and distribution–based measures

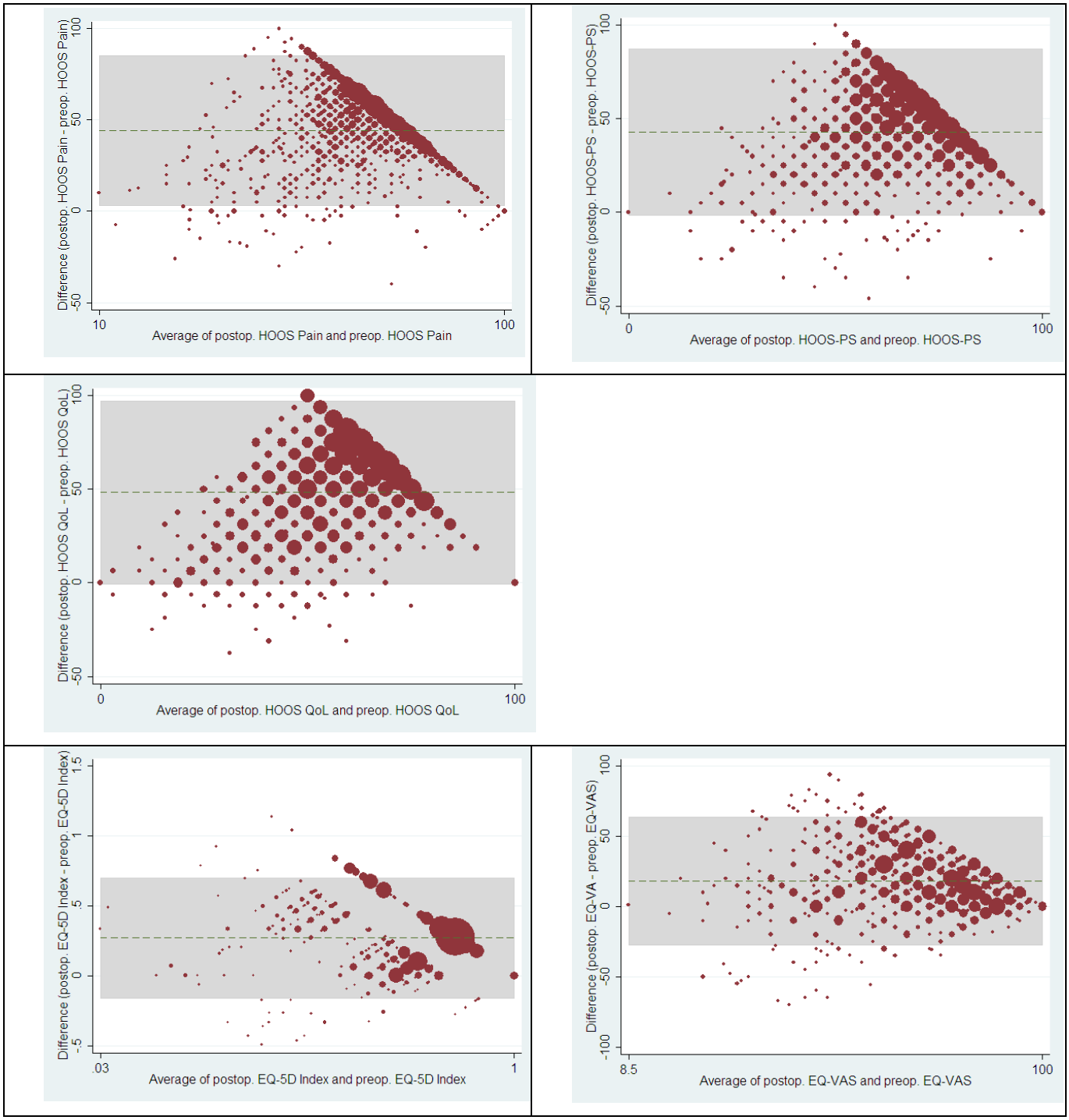
The MCII reported by the primary approach (Paper IV, Table 2) was very similar to the SD of change for all PROs, and our results are in contrast with the results of Norman et al., who found the MCII to be approximately ½ SD (142). Several papers questioning this approximation has been published (220-222). The MCII reported by the mean change approach is situated within the limits of agreement for all PROs which indicate that the estimated cut-points are over the level of measurements error. The Bland-Altman plots in Table 26 and Table 27 are distinctively different as the first represent the change in preoperative- to postoperative status and the latter represent the test-retest item- and sum score agreement. The standard error of the mean express how reliable an estimate of the mean is. The standard error of the mean found was small for all PROs compared to the MCII estimated, due to the sample size (Table 28).

When the MCII exceed the MDC, it implies that the MCII is true, and not only a measurement error (144;162). The change reflected in the MDC is assumed to be the same across the range of possible scores, but since this is often not the case, the MDC should only be considered a guideline (144). For HOOS QoL the MDC was higher than the MCII, and for EQ-VAS the MDC and MCII were the same, implying that the measurement errors for these PROs are bigger or equal to the estimated MCII, and suggest a suboptimal interpretability of change for these subscales.

All ESs were large, and all PROs had a good sensitivity for detecting clinical changes. ES is a 'signal-to-noise ratio', and since it has no units, different PROs may be directly compared in terms of the variability among individuals (132). ES does not take into account the variability of change and has the limitation that it is strongly influenced by the level of heterogeneity of the sample: a small baseline SD gives a larger ES (161). Our finding of an ES of 1.4 (Table 28) for the EQ-5D Index corresponds well with previously reported ES of EQ-5D Index findings of 1.3 (109). The SRM were large for all PROs, as might be expected for THA patients (153), which indicate that the change in PRO scores is large compared to the background variability (161). I found the same SRM as previously reported (2.1) for the HOOS Pain, the SRM of HOOS QoL was similar (1,9 vs. 1.6) and the SRM of HOOS-PS (1.9) was similar to previously reported SRM for HOOS subscale Function in Daily Living (1.3) and HOOS subscale Sport and Recreation Function (1.7) (Table 28) (223). The reported SRM for EQ-5D Index and EQ-VAS was higher than the SRM reported in RA patients at 3 months of follow-up (218), probably due the different interventions. ES and SRM are independent of sample size (161).

The SEM was considerably lower than the reported MCII for all PROs, and the MCIIs were beyond the limits of ±1.96 SEM indicating that the observed changes were likely to reflect true change, and not an artifact of measurement error (132). Several authors have found the SEM to be close to the MCII (147;157;224), but that the MCII should be approximately 1 SEM is not any more meaningful than any other value (154). The SEM reported in Table 28, can be explained by the high baseline values in study IV (225). Our finding of a SEM of 0.08 (Table 28) for the EQ-5D Index corresponds well with previous findings of a SEM of 0.12 for THA patients 6 months after surgery (109). RCI is a statistic that determines the magnitude of change score necessary of a given PRO to be considered statistically reliable and represents the number of scale points needed on a PRO to determine if a change in score is due to real change or chance variation (226). If RCI is higher than the cut-point (1.96) the change is statistically significant (161;162). Only the EQ-VAS had a

Table 26. Additional results: Bland-Altman plots, from Study IV

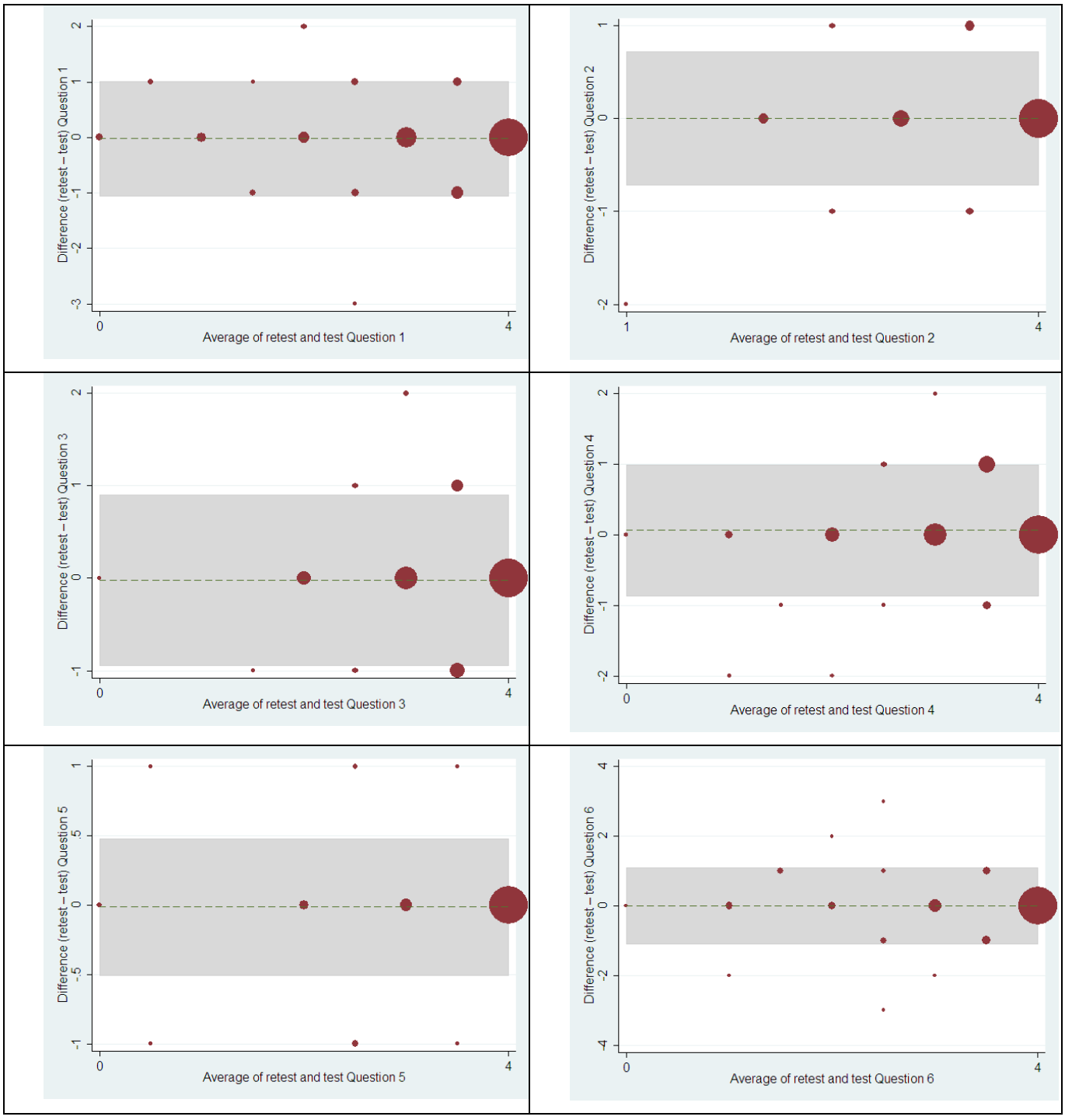


RCI below 1.96 (Table 28). This may partly be explained by that the RCI approach is more conservative than the SEM approach (161). SEM and RCI quantifies the amount of error inherent in the PRO and the amount of random variation that can be expected in repeated administrations, is quite unaffected by sample size, is less influenced by variability in the sample than ES, and is less influenced by the variability of the change than SRM (161).

The PASS cut-points reported by the primary approach (Paper IV, Table 3) were higher than the sample mean, but not higher than the median, and all PASS cut-points were within the inter quartile range (Table 29). All in all, the distribution based reliability measures contribute to the validity of the MCII and PASS estimations.

The distribution based reliability measures for the hip-specific OHS in study III (Table 5) is very different from the distribution based reliability measures of the hip-specific HOOS in study IV. The mean change score of OHS was 0.05 (OHS range from 0-48), whereas the change score for the HOOS Pain was 44, the change score for the HOOS-PS was 43 and the change score for the HOOS QoL was 48 (all HOOS subscales range from 0-100). This difference is easily explained: the test-retest of OHS included 166 patients who answered OHS twice within two weeks in a steady

Table 27. Additional results: Bland-Altman plot for OHS, from Study III

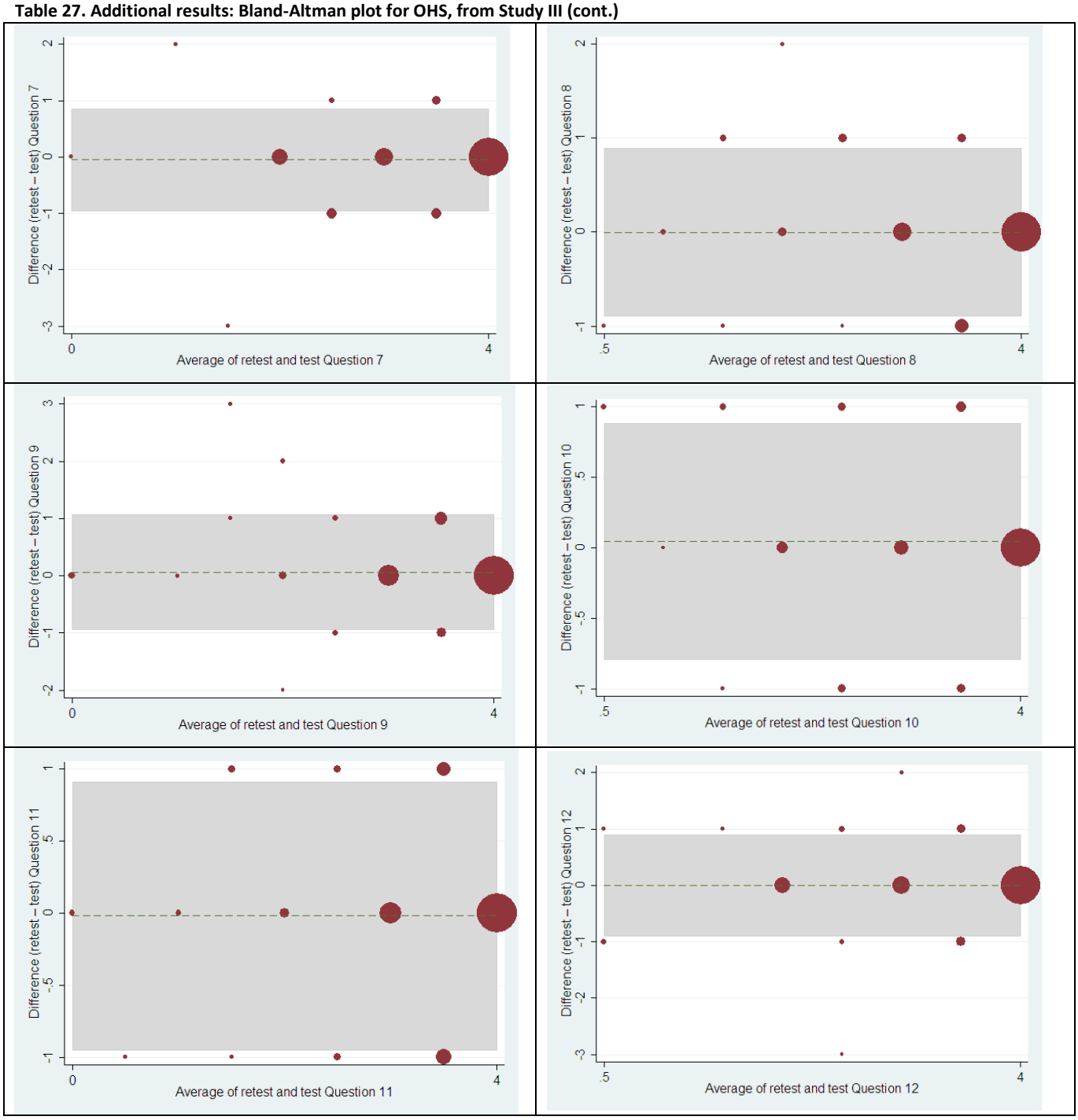


state after minimum three years postoperatively. In study IV, 1,239 patients were included, HOOS and EQ-5D where not answered in a steady state, but answered preoperative and postoperative. The difference in the OHS test-retest was expected to be very small, and the differences from preoperative to postoperative for HOOS and EQ-5D were expected to be large, and a comparisons of the distribution based reliability measures of OHS versus HOOS or EQ-5D, would not be meaningful.

PRO criticism

There are several potential problematic aspects of using different PROs. Giesinger et al. found a strong relationship between psychological status and orthopaedic outcome for WOMAC and the Forgotten Joint Score-12, indicating poor divergent validity. Their findings suggest that these PROs may not adequately reflect the category names of the constructs assessed (pain, stiffness, function or joint awareness), but also to a high degree reflect the patients psychological status (227). Baumann et al. reported strong associations between immediate postoperative patient

Table 27. Additional results: Bland-Altman plot for OHS, from Study III (cont.)

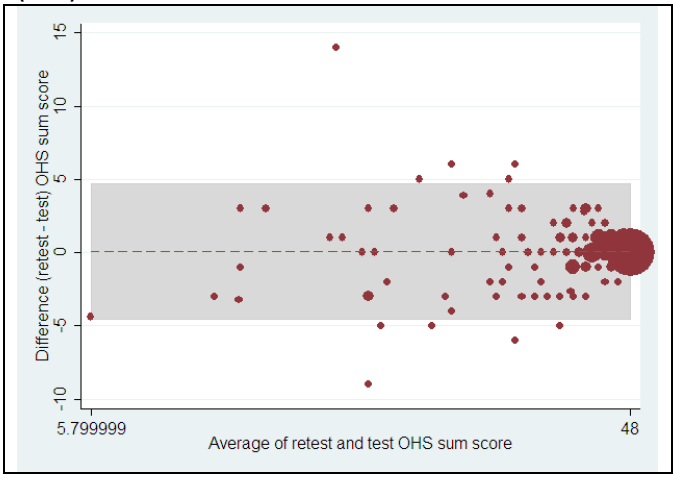


satisfaction with care and 1 year postoperative PRO scores for the dimensions 'bodily pain', 'mental health', 'social functioning', 'vitality' and 'general health' of the SF-36, suggesting that these dimensions may be affected by patient satisfaction (71). PROs have been reported to be less sensitive to deterioration in functional status with advancing age than performance-based measures, as patients seem to tolerate more functional limitations and adapt to a certain amount of declining function when they get older (228). The small to medium correlations reported between PROs and performance-based measures(62) point towards that PROs and performance-based measures are complementary, but may not seem to measure exactly the same constructs (228).

STRENGTHS AND LIMITATIONS

Several methodological problems must be considered when interpreting our results. I have not validated the measurement properties of the Danish language versions for all PROs included in the studies. These PROs has been validated by others, and all included PROs are often used and considered well validated. The EQ-5D was used in all studies, and the EQ-5D Index had a bi-modal distribution in our data, as previously reported by others (229), probably due to the EQ-5D algorithm. The implication is that the uncertainties of the results are greater than described by the CI and p-values, and all consequences of this may not yet be known, which has to be taken into consideration when interpreting our results.

Table 27. Additional results: Bland-Altman plot for OHS, from Study III (cont.)



I have not considered potential methodological difficulties in the development of the EQ-5D health state valuation (171).

The postoperative sum scores in study I were not normal distributed, but due to the group size (2,365-2,419 patients), 95% CI and normal based methodology was used (Paper I, Table 2), on basis on the central limit theorem. The results are valid, but the mean may not be the best summary of the distribution, and the median scores, IQRs and ranges have been included.

prosthesis types (HOOS Pain, HOOS-PS, OHS, SF-12 MCS, EQ-VAS, $p < 0.001$-0.05), and different age groups (HOOS-PS, SF-12 MCS, EQ-VAS, $p < 0.001$-0.04). In PROs with unequal variances the maximal variance ratios were highest for different diagnoses (1.4-1.7), compared to different gender (1.0-1.1), different prosthesis types (1.0-1.1) and different age groups (1.2). This may complicate the interpretation of discriminative ability (Paper I, Table 4). No information concerning revision surgery or subsequent contra lateral THA was available.

Study III is a secondary data analysis and I have solely included postoperative patients. The psychometric properties of PROs used in elective surgical contexts are usually largely evaluated on pre-operative data, making the interpretation of our ceiling effect, skew and internal consistency more demanding. Since the patients are all postoperative I expected the OHS to be highly skewed, and it could therefore be argued that speaking of ceiling effects would be misleading. I argue that it is important to assess postoperative development, and have chosen to report the percentage of ceiling at PRO level, even though this characteristic would more often be assessed at the individual item level in PROs development. Further studies on the responsiveness and sensitivity to the Danish version of the OHS are warranted. Patients who received two disease-specific PROs answered the HOOS a median of 4.9 years postoperatively (range 0.9-10.5 years) and the OHS a median of 7.1 years (range 3.1-12.8 years) postoperatively, when both PROs presumably measured the patient's health status during a period in which their hip function

Table 28. Additional results: Distribution based measures of change in HOOS and EQ-5D, from Study IV *1*

PROs	Mean change score	Standard deviation (SDchange)	Limits of agreement (LOA)	Standard error of measurement (SEM)	Effect size (ES)	Minimal detectable change (MDC)	Standardized response mean (SRM)	Standard error of the mean	Reliability change index (RCI)
HOOS Pain	44	21	4-85	5	2.7	15	2.1	0.61	5.77
HOOS-PS	43	23	-2-87	7	2.4	18	1.9	0.66	4.57
HOOS QoL	48	25	0-97	8	3.0	21	1.9	0.65	4.56
EQ-5D Index	0.27	0.22	-0.16-0.70	0.08	1.4	0.23	1.2	0.09	2.31
EQ-VAS	18	23	-28-64	8	0.9	23	0.8	0.67	1.53

1: All patients included

Table 29. Additional results: Mean, Median, Standard deviation and Interquartile range of postoperative scores in HOOS and EQ-5D, from Study IV *1*

PROs	Mean	Median	Standard deviation (SD)	Interquartile range (IQR)
HOOS Pain	89	97	16	83-100
HOOS-PS	85	90	18	75-100
HOOS QoL	80	88	22	69-100
EQ-5D Index	0.88	1.00	0.16	0.78-1.00
EQ-VAS	80	85	18	75-90

1: All patients included

Logistic regression was used to compare overall feasibility criteria between different PROs, adjusting for age, sex, primary hip diagnosis and prosthesis type. Due to small group sizes for floor effect (n = 0-13), percentile based CI may have been more appropriate (Paper I, Table 3). Variance tests revealed unequal variances in some PROs for different diagnoses (HOOS Pain, HOOS-PS, HOOS QoL, EQ-VAS, $p < 0.001$-0.01), different gender (HOOS Pain, HOOS-PS, HOOS QoL, OHS, $p < 0.001$), different

was in the same steady state. I did not exclude patients who had undergone revision surgery, or received contra lateral THA following the index operation. No information concerning rehabilitation programs or postoperatively occurring co-morbidity of the patients were available, and I cannot exclude the possibility of that these factors may have affected the PRO scores.

In study IV, I have not validated the measurement properties of the included additional questions. These additional questions have been used in other studies, some with minor modifications

(167;168;230-234). Four questions (regarding previous joint replacements, postoperative physical therapy, knee symptoms and back symptoms) were developed in cooperation with clinical experts, but have not been used in the studies reported in this thesis. I translated two of the anchor items from English without a formal cross-cultural validation. The EQ-5D Index had an anchor-change score correlation of less than 0.30 for both the hip specific change anchor and the general health change anchor. EQ-VAS had an anchor-change score correlation of less than 0.30 for the hip specific change anchor. All other MCII estimations were based on moderate (<0.50) anchor-change score correlations. A retrospective transition anchor was used for MCII estimations of HOOS and an absolute change anchor was used for MCII estimations of EQ-5D. The sex ratio and mean age for each sex group were very similar between our study population and the entire Danish 2010 THA population, whereas the distribution of diagnoses were somewhat different; a higher percentage of the patients in our study had idiopathic OA and other arthritis (26). Patients who declined to participate in the study were slightly older and more often operated due to childhood hip diseases than included patients, possible producing bias in relation to age and diagnoses estimates.

In study I, our results have high external validity since the distribution of age groups, the sex ratio, diagnoses, and types of prosthesis were similar between our study population and the entire Danish THA population, as well as hip replacement populations seen in other hip registries. Regarding knee arthroplasty, Dunbar (2001) compared properties of the SF-12 and the Oxford knee score in a knee registry setting and found response rates, percentages of fully completed questionnaires, and floor and ceiling effects comparable with our findings from the SF-12 and OHS, suggesting generalizability of our results (4). I minimized selection bias by randomly selecting patients for inclusion and checked for equal age and sex composition in the groups.

I believe study II is representative of a wide variety of research and clinical settings where paper form questionnaires are used. THA is indicated for patients with pain and functional disabilities or reduced quality of life. The population is an extensively studied elderly population, with a mean age in Denmark of 70/67 years (female/male), the patients have a spectrum of comorbid conditions and they constitute a suitable and interesting population in relation to validation of AFP.

I had an excellent response rate in study I and III. In study III, I included a range of patients from 30 to 80 years. Most patients get their THA in this age range. The study III population is slightly younger than the Danish THA population, but I believe that our results have high external validity since the gender ratio and diagnoses are similar between the study population and the Danish THA population. The Danish OHS was validated in the context of a THA registry, compared with both generic and disease-specific PROs and examined 1-2, 5-6 and 10-11 years following THA.

I consider the results of study IV to have high external validity due to the inclusion of approximately 15% of the entire Danish THA production of 2010 from 16 centers dispersed all over Denmark, both centers with low and high productions, public as well as private, and both university hospitals and community hospitals.

CONCLUSIONS

PAPER I

The HOOS, the OHS, the SF-12, and the EQ-5D are all appropriate PROs for administration in a hip registry. I found minor differences between the disease-specific and the generic PROs regarding ceiling and floor effects as well as discarded items. Group sizes from 51 to 1,566, depending on descriptive factors and choice of PRO, were needed for subgroup analysis.

PAPER II

AFP can yield excellent results provided use of highly structured questionnaires. OMR performed equally as well as manual double-key entering, and better than single-key entering. Regarding ICR, I cannot draw firm conclusions due to the limited data available in this study, and therefore further research, as well as improvement in ICR technology, is warranted.

PAPER III

The Danish version of the OHS had good feasibility, an excellent response rate, no floor effect, but a high ceiling effect as was expected with our post-operative patients and few patients missed too many items to calculate a sum score. The Danish version of the OHS is a valid and reliable tool for outcome studies on THR patients, in comparison with the HOOS, EQ-5D and SF-12, and can be used in a hip registry setting.

PAPER IV

Using a population-based cohort design, we determined cut-points for the change representing the MCII and for the postoperative score representing the PASS 1 year after THA for HOOS Pain, HOOS-PS, HOOS QoL, EQ-5D Index, and EQ-VAS. This study facilitates interpretability of PRO scores and may improve understanding of PRO findings in future THA outcome studies. MCIIs corresponded to a 38–55% improvement from mean baseline PRO score and PASSs corresponded to absolute follow-up scores of 57–91% of the maximum score in THA patients 1 year after surgery, which may serve as reference values in registry settings.

FUTURE PERSPECTIVES

The studies presented can give rise to many future studies; I have two large cohorts of patients and it would be very interesting to follow these cohorts to examine the effect of THA on pain, physical function and long term quality of life.

In study IV, I included a substantial set of items besides the PROs. The preoperative questionnaire included items regarding height, weight, marital status, education, previous joint replacements, knee symptoms, back symptoms, general health, diet, tobacco and alcohol consumption, medication, degree of physically demanding occupation, general physical strain, and physical activity. The postoperative questionnaire included items regarding weight, marital status, general health, back symptoms, knee symptoms, patient reported operation result, patient reported hip improvement, medication, general physical strain, physical activity, and physical therapy.

Examining the items not yet examined in the presented studies and including PROs in hip arthroplasty registries would lead to a much better overview over the patients' perspective, and could lead to improvements in the treatment of THA patients in several ways;

Identifying patients at risk preoperatively

10-15% of patients report persistent pain and functional limitation postoperatively (43), and 14-36% of patients do report that they have not benefitted from the operation (44). In study IV, I found some patients with very little hip improvement after THA (Paper IV, Table 2), and identifying these patients at risk preoperatively could affect decision making and indication for surgery, and would enable surgeons to improve preoperative information as well as to tailor the postoperative interventions in order to improve the outcome. Besides better postoperative outcome, this could lead to cutting of economic costs, as well as removing operation-related risks and disadvantages for patients not benefitting from the operation.

Identifying patients at risk postoperatively

Many resources are used for follow-up of THA patients postoperatively. They often include radiological examination and clinical examination, and are time- and cost demanding. The implant survivals following THA are very high, and most patients also have a subjective successful outcome. By using PROs and anchor questions in the postoperative course, it could be possible to reduce the follow-up of patients with successful THA, and focus on the patients at risk. Patients with low PRO scores postoperative or patients not benefitting from the operation (low change scores), could be scheduled for additional follow-ups and have a more closely monitored course. This could lead to a better result for these patients, and could also possibly be cost-reducing.

Identifying inferior implants and inferior surgery approaches

By including PROs in hip arthroplasty registries, it could be possible to identify inferior (or superior) implants, fixation methods and surgery approaches that are not possible to identify based on hard endpoints. It is likely that different implants or surgery approaches may give different results in terms of pain, physical function and quality of life. Recent data suggest that a posterior approach may give better satisfaction and less pain than a direct lateral approach (19). By choosing implants, fixation methods and surgery approaches that show better patient reported results in hip arthroplasty registries, the general outcome after THA may be improved.

SUMMARY

PROs are used increasingly in orthopedics and in joint registries, but still many aspects of use in this area have not been examined in depth. To be able to introduce PROs in the DHR in a scientific fashion, my studies were warranted;

The feasibility of four often used PROs (OHS, HOOS, EQ-5D and SF-12) was examined in a registry context. Having the PROs in the target language is an absolute necessity, so I translated, cross-culturally adapted and validated a Danish language version of an often used PRO (OHS), since this PRO had no properly developed Danish language version. To minimize data loss and to maximize the data quality I validated our data capture procedure; an up to date AFP system, by comparing scannable, paper-based PROs, with manual single-key- and double-key entered data. To help further registry-PRO studies, I calculated the number of patients needed to discriminate between subgroups of age, sex, diagnosis, and prosthesis type for each of four often used PROs (OHS, HOOS, EQ-5D and SF-12), and to simplify the clinical interpretation of PRO scores and PRO change scores in PRO studies, I estimated MCII and PASS for two often used PROs (EQ-5D and HOOS).

The feasibility study included 5,747 THA patients registered in the DHR, and I found only minor differences between the disease-specific and the generic PROs regarding ceiling and floor effects as well as discarded items. The HOOS, the OHS, the SF-12, and the EQ-5D are all appropriate PROs for administration in a hip registry. I found that group sizes from 51 to 1,566 were needed for subgroup analysis, depending on descriptive factors and choice of PRO.

The AFP study included 200 THA patients (398 PROs, 4,875 items and 21,887 data fields), and gave excellent results provided use of highly structured questionnaires. OMR performed equally as well as manual double-key entering, and better than single-key entering.

The PRO translation and validation study included 2,278 patients (and 212 patients for the test-retest). I found that the translated PRO had good feasibility, an excellent response rate, no floor effect, but a high ceiling effect (as was expected with our postoperative patients) and few patients missed too many items to calculate a sum score. The translated PRO had high test-retest reliability and very high internal consistency, and appears to be a valid and reliable tool for outcome studies on THA patients in a hip registry setting.

The MCII and PASS study included 1,335 patients, and I estimated that one year after THA, an improvement of 38-55% from mean baseline PRO score and absolute follow-up scores of 57-91% of the maximum score correspond to a minimal important improvement and acceptable symptom state, respectively.

Table 30. Abbreviations

AFP	**Automated Form Processing**
AUC	**Area under the curve**
CI	**Confidence intervals**
DHR	**The Danish Hip Arthroplasty Registry**
EQ-VAS	**The visual analogue scale part of EQ-5D**
EQ-5D	**EuroQoL-5D-3L**
EQ-5D Index	**A global health index with a weighted total value for HRQoL**
HOOS	**Hip dysfunction and Osteoarthritis Outcome Score**
HOOS Pain	**HOOS subscale pain**
HOOS PS	**HOOS-Physical Function Shortform**
HOOS QoL	**HOOS subscale hip-related quality of life**
HR	**Health-related**
ICR	**Intelligent Character Recognition**
MCII	**Minimal Clinically Important Improvement**
MCS	**Mental Component Summary of SF-12**
OA	**Osteoarthritis**
OHS	**Oxford Hip Score**
OMR	**Optic Mark Recognition**
OR	**Odds ratio**
PASS	**Patient Acceptable Symptom State**
PCS	**Physical Component Summary of SF-12**
PRO	**Patient Reported Outcome measure; a questionnaire (abbr. 'PROM' used in study III)**
QoL	**Quality of life**
RA	**Rheumatoid arthritis**
ROC	**Receiver operating characteristic**
SF-12	**SF-12 Health Survey**
THA	**Total Hip Arthroplasty (abbr. 'THR' (Total Hip Replacement) used in study III)**
VAS	**Visual Analog Scales**

Table 31. List of terms and definitions

Ceiling effect	**Percentage of the sample achieving the best possible scores (1)**
Construct	**A well-defined and precisely demarcated subject of measurement (by psychologists used for unobservable characteristics, such as 'health-related quality of life') (2)**
Construct validity	**The degree to which the scores of a PRO instrument are consistent with hypotheses (for instance with regard to internal relationships, relationships to scores of other instruments, or differences between relevant groups) based on the assumption that the PRO instrument validly measures the construct to be measured (3)**
Content validity	**The degree to which the content of an HR-PRO instrument is an adequate reflection of the construct to be measured (3)**
Domain	**A sub-score within a questionnaire meant to cover a specific condition of interest, e.g. 'Bodily Pain', which is a domain within the SF-12 (4). In this thesis used interchangeably with 'dimension' and 'subscale'**
Feasibility	**The usability of a questionnaire in a specific setting, including response rate, floor- and ceiling effect, missing items and need for manual validation (5)**
Floor effect	**Percentage of the sample achieving the worst possible scores (1)**
Internal consistency	**The degree of the interrelatedness among the items (3)**
Item	**A single question within a domain or questionnaire (4). In this thesis used interchangeably with 'question'**
Likert scale	**A rating scale in which raters express their opinion on a given subject by marking a box within a continuum of disagree-agree statements (4)**
Manual validation	**Validation of code for the questionnaire answer in question by a human operator when an automated forms processing system cannot convert an answer due to poor or ambiguous questionnaire completion (6)**
Reliability	**The extent to which scores for patients who have not changed are the same for repeated measurement under several conditions: for example, using different sets of items from the same PROs (internal consistency), over time (test-retest) by different persons on the same occasion (inter rater) or by the same persons (i.e., raters or responders) on different occasions (intra rater) (3)**
Response rate	**The proportion of respondents in relation to all patients who received the questionnaire (7)**
Validity	**The degree to which a PRO instrument measures the construct(s) it purports to measure (3)**

REFERENCES

1 McHorney CA, Tarlov AR. Individual-Patient Monitoring in Clinical Practice: Are Available Health Status Surveys Adequate? Qual Life Res. 1995 Aug 1;4(4):293-307.

2 de Vet HC, Terwee CB, Mokkink L, Knol DL. Concepts, theories and models, and types of measurements. Measurements in medicine-Practical guides to biostatistics and epidemiology.Cambridge: Cambridge University Press; 2011. p. 7-29.

3 Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. 2010 Jul 1;63(7):737-45.

4 Dunbar MJ. Subjective outcomes after knee arthroplasty. Acta Orthop Scand Suppl. 2001 Feb;72(301):1-63.

5 Paulsen A, Pedersen AB, Overgaard S, Roos EM. Feasibility of 4 patient-reported outcome measures in a registry setting. Acta Orthopaedica. 2012 Aug 1;83(4):321-7. doi: 10.3109/17453674.2012.702390.

6 Paulsen A, Overgaard S, Lauritsen JM. Quality of Data Entry Using Single Entry, Double Entry and Automated Forms Processing - An Example Based on a Study of Patient-Reported Outcomes. PLoS ONE. 2012 Apr 6;7(4):e35087. doi: 10.1371/journal.pone.0035087.

7 Rolfson O. Patient-reported Outcome Measures and Health-economic Aspects of Total Hip Arthroplasty -A study of the Swedish Hip Arthroplasty Register Institute of Clinical Sciences at Sahlgrenska Academy, University of Gothenburg; 2010.

8 MacLennan WJ. History of arthritis and bone rarefaction evidence from paleopathology onwards. Scott Med J. 1999 Feb;44(1):18-20.

9 Gomez PF, Morcuende JA. Early attempts at hip arthroplasty--1700s to 1950s. Iowa Orthop J. 2005;25:25-9.

10 Berry DJ, Harmsen WS, Cabanela ME, Morrey BF. Twenty-five-Year Survivorship of Two Thousand Consecutive Primary Charnley Total Hip Replacements Factors Affecting Survivorship of Acetabular and Femoral Components. The Journal of Bone & Joint Surgery. 2002 Feb 1;84(2):171-7.

11 Callaghan JJ, Albright JC, Goetz DD, Olejniczak JP, Johnston RC. Charnley Total Hip Arthroplasty with Cement Minimum Twenty-five-Year Follow-up*. The Journal of Bone & Joint Surgery. 2000 Apr 1;82(4):487.

12 Overgaard S, Pedersen AB. Dansk Hoftealloplastik Register Årsrapport 2009. 2010.

13 Engesaeter LB, Furnes O, Havelin LI, Fenstad A.M. Nasjonalt Register for Leddproteser RAPPORT Juni 2010. 2010.

14 Porter M, Borroff M, Gregg P, Howard P, MacGregor A, Tucker K. National Joint Registry for England and Wales 7th Annual Report 2010. 2011.

15 Australian Orthopaedic Association. Australian Orthopaedic Association National Joint Replacement Registry. Annual Report 2010. Adelaide; 2010.

16 Canadian Institute for Health Information. Hip and Knee Replacements in Canada -Canadian Joint Replacement Registry (CJRR) 2008-2009 Annual Report. Ottawa, Ont.: CIHI; 2009.

17 New Zealand Orthopaedic Association. The New Zealand Joint Registry. Eleven Year Report. January 1999 to December 2009. 2010.

18 NHS National Services Scotland. Scottish Arthroplasty Project Report 2010. 2010.

19 Garellick G, Kärrholm J., Rogmark C., Herberts P. Svenska Höftprotesregistret Årsrapport 2011. 2012.

20 Learmonth ID, Young C, Rorabeck C. The operation of the century: total hip replacement. The Lancet. 2007 Oct 27;370(9597):1508-19. doi: 10.1016/S0140-6736(07)60457-7.

21 Porter M, Borroff M, Gregg P, MacGregor A, Tucker K. National Joint Registry for England and Wales 6th Annual Report 2009. 2010.

22 Garellick G, Kärrholm J, Rogmark C., Herberts P. Swedish Hip Arthroplasty Register Annual Report 2008 Shortened Version. Department of Orthopaedics, Sahlgrenska University Hospital; 2010.

23 Engesaeter LB, Furnes O, Havelin LI, Fenstad A.M. Nasjonalt Register for Leddproteser RAPPORT Juni 2010. 2010.

24 Liang MH, Cullen KE, Poss R. Primary total hip or knee replacement: evaluation of patients. Ann Intern Med. 1982 Nov;97(5):735-9.

25 Engesaeter LB, Furnes O, Havelin LI. Nasjonalt Register for Leddproteser RAPPORT Oktober 2012. 2012.

26 Overgaard S. Dansk Hoftealloplastik Register Årsrapport 2012. 2012.

27 American Academy of Orthopaedic Surgeons. OrthoInfo -Total Hip Replacement. 2012 [cited 2012 Dec 4]. [Internet]. Available from: http://orthoinfo.aaos.org/topic.cfm?topic=A00377.

28 Powers-Freeling L. National Joint Registry for England and Wales 9th Annual Report 2012. 2012.

29 Badley EM, Crotty M. An international comparison of the estimated effect of the aging of the population on the major cause of disablement, musculoskeletal disorders. J Rheumatol. 1995 Oct;22(10):1934-40.

30 Pedersen AB, Johnsen SP, Overgaard S, Soballe K, Sorensen HT, Lucht U. Total hip arthroplasty in Denmark: incidence of primary operations and revisions during 1996-2002 and estimated future demands. Acta Orthop. 2005 Apr;76(2):182-9.

31 Birrell F, Johnell O, Silman A. Projecting the need for hip replacement over the next three decades: influence of changing demography and threshold for surgery. Ann Rheum Dis. 1999 Sep;58(9):569-72.

32 Pedersen AB, Johnsen SP, Overgaard S, Soballe K, Sorensen HT, Lucht U. Total hip arthroplasty in Denmark: incidence of primary operations and revisions during 1996-2002 and estimated future demands. Acta Orthop. 2005 Apr;76(2):182-9.

33 Paavolainen P, Hamalainen M, Mustonen H, Slatis P. Registration of arthroplasties in Finland. Acta Orthopaedica. 1991 Jan 1;62(s241):27-30. doi: 10.3109/17453679109155101.

34 Havelin LI, Engesaeter LB, Espehaug B, Furnes O, Lie SA, Vollset SE. The Norwegian Arthroplasty Register: 11 years and 73,000 arthroplasties. Acta Orthopaedica. 2000;71(4):337-53.

35 Havelin LIM, Robertsson OM, Fenstad AMM, Overgaard SM, Garellick GM, Furnes OM. A Scandinavian Experience of Register Collaboration: The Nordic Arthroplasty Register Association (NARA). [Article]. Journal of Bone & Joint Surgery - American Volume. 2011 Dec;93 Suppl 3:13-9.

36 Graves SE. The value of arthroplasty registry data. Acta Orthopaedica. 2000;0(0):1-2.

37 Vincent D, Marx C, Rankin EA, Batten JC, Frank CB, Atkinson D, et al. Position Statement in Support of National Joint Registries. The Journal of Bone & Joint Surgery. 2009 Dec 1;91(12):2983. doi: 10.2106/JBJS.I.01469.

38 Britton AR, Murray DW, Bulstrode CJ, McPherson K, Denham RA. Pain levels after total hip replacement: their use as endpoints for survival analysis. J Bone Joint Surg Br. 1997 Jan;79(1):93-8.

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

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

41 Soderman P, Malchau H, Herberts P, Zugner R, Regner H, Garellick G. Outcome after total hip arthroplasty: Part II. Disease-specific follow-up and the Swedish National Total Hip Arthroplasty Register. Acta Orthop Scand. 2001 Apr;72(2):113-9.

42 Lucht U, Johnsen SP. Dansk Hofteallоplastik Register Årsrapport 2005. 2005.

43 Nikolajsen L, Brandsborg B, Lucht U, Jensen TS, Kehlet H. Chronic pain following total hip arthroplasty: a nationwide questionnaire study. Acta Anaesthesiol Scand. 2006;50(4):495-500. doi: 10.1111/j.1399-6576.2006.00976.x.

44 Judge A, Cooper C, Williams S, Dreinhoefer K, Dieppe P. Patient-reported outcomes one year after primary hip replacement in a European Collaborative Cohort. Arthritis Care Res. 2010;62(4):480-8. doi: 10.1002/acr.20038.

45 Harris WH. Traumatic arthritis of the hip after dislocation and acetabular fractures: treatment by mold arthroplasty. An end-result study using a new method of result evaluation. 1969 Jun.

46 Charnley J. The long-term results of low-friction arthroplasty of the hip performed as a primary intervention. J Bone Joint Surg Br. 1972 Feb;54(1):61-76.

47 Janse AJ, Gemke RJBJ, Uiterwaal CSPM, van der Tweel I, Kimpen JLL, Sinnema G. Quality of life: patients and doctors don't always agree: a meta-analysis. J Clin Epidemiol. 2004 Jul;57(7):653-61. doi: 10.1016/j.jclinepi.2003.11.013.

48 Givon U, Ginsberg GM, Horoszowski H, Shemer J. Cost-utility analysis of total hip arthroplasties. Technology assessment of surgical procedures by mailed questionnaires. Int J Technol Assess Health Care. 1998;14(4):735-42.

49 O'Boyle CA. Assessment of quality of life in surgery. Br J Surg. 1992 May;79(5):395-8.

50 Herberts P, Malchau H. How outcome studies have changed total hip arthroplasty practices in Sweden. Clin Orthop Relat Res. 1997 Nov;(344):44-60.

51 Malchau H, Garellick G, Eisler T, Karrholm J, Herberts P. Presidential guest address: the Swedish Hip Registry: increasing the sensitivity by patient outcome data. Clin Orthop Relat Res. 2005 Dec;441:19-29.

52 Wylde V, Blom AW. The failure of survivorship. [Editorial]. Journal of Bone & Joint Surgery - British Volume. 2011 May;93B(5):569-70.

53 US Department of Health and Human Services (USDHHS). Draft guidance for industry. Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. 2006 [Internet]. Available from: www.ispor.org/workpaper/FDAPROGuidance2006.pdf.

54 US Department of Health and Human Services (USDHHS). Guidance for industry. Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. 2009 [Internet]. Available from: www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM193282.pdf.

55 Rolfson O. Patient-reported Outcome Measures and Health-economic Aspects of Total Hip Arthroplasty -A study of the Swdish Hip Arthroplasty Register Institute of Clinical Sciences at Sahlgrenska Academy, University of Gothenburg; 2010.

56 Rothwell AG, Hooper GJ, Hobbs A, Frampton CM. An analysis of the Oxford hip and knee scores and their relationship to early joint revision in the New Zealand Joint Registry. J Bone Joint Surg Br. 2010 Mar;92(3):413-8.

57 Rolfson OM, Rothwell AC, Sedrakyan AM, Chenok KEM, Bohm EBMMF, Bozic KJM, et al. Use of Patient-Reported Outcomes in the Context of Different Levels of Data *. [Article]. Journal of Bone & Joint Surgery - American Volume. 2011 Dec;93 Suppl 3:66-71.

58 Horan FT. Joint registries. [Editorial]. Journal of Bone & Joint Surgery - British Volume. 2010 Jun;92(6):749-50.

59 Devlin NJ, Parkin D, Browne J. Patient-reported outcome measures in the NHS: new methods for analysing and reporting EQ-5D data. Health Econ. 2010 Aug;19(8):886-905.

60 Daubney ME, Culham EG. Lower-extremity muscle force and balance performance in adults aged 65 years and older. Phys Ther. 1999 Dec;79(12):1177-85.

61 Nussbaumer S, Leunig M, Glatthorn J, Stauffacher S, Gerber H, Maffiuletti N. Validity and test-retest reliability of manual goniometers for measuring passive hip range of motion in femoroacetabular impingement patients. BMC Musculoskeletal Disorders. 2010;11(1):194. doi: 10.1186/1471-2474-11-194.

62 Farag I, Sherrington C, Kamper SJ, Ferreira M, Moseley AM, Lord SR, et al. Measures of physical functioning after hip fracture: construct validity and responsiveness of performance-based and self-reported measures. Age Ageing. 2012 Sep 1;41(5):659-64.

63 Latham NK, Mehta V, Nguyen AM, Jette AM, Olarsch S, Papanicolaou D, et al. Performance-based or self-report measures of physical function: which should be used in clinical trials of hip fracture patients? Arch Phys Med Rehabil. 2008 Nov;89(11):2146-55.

64 Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007 Jan;60(1):34-42.

65 Mokkink L, Terwee C, Patrick D, Alonso J, Stratford P, Knol D, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res. 2010 May 1;19(4):539-49.

66 Wild D, Grove A, Martin M, Eremenco S, McElroy S, Verjee-Lorenz A, et al. Principles of Good Practice for the Translation and Cultural Adaptation Process for Patient-Reported Outcomes (PRO) Measures: Report of the ISPOR Task Force for Translation and Cultural Adaptation. Value in Health. 2005;8(2):94-104. doi: 10.1111/j.1524-4733.2005.04054.x.

67 Lawlor DA, Chaturvedi N. Methods of measurements in epidemiology - call for a new type of paper in the IJE. Int J Epidemiol. 2010 Oct 1;39(5):1133-6.

68 McLeod LD, Coon CD, Martin SA, Fehnel SE, Hays RD. Interpreting patient-reported outcome results: US FDA guidance and emerging methods. Expert Rev Pharmacoeconomics Outcomes Res. 2011 Apr 1;11(2):163-9. doi: 10.1586/erp.11.12.

69 Wylde V, Hewlett S, Learmonth ID, Cavendish VJ. Personal impact of disability in osteoarthritis: patient, professional and public values. Musculoskelet Care. 2006;4(3):152-66. doi: 10.1002/msc.86.

70 Wright JG, Rudicel S, Feinstein AR. Ask patients what they want. Evaluation of individual complaints before total hip replacement. J Bone Joint Surg Br. 1994 Mar 1;76-B(2):229-34.

71 Baumann C, Rat AC, Osnowycz G, Mainard D, Cuny C, Guillemin F. Satisfaction with care after total hip or knee replacement predicts self-perceived health status after surgery. BMC Musculoskelet Disord. 2009;10:150.

72 Strauss ME, Smith GT. Construct Validity: Advances in Theory and Methodology. Annu Rev Clin Psychol. 2009 Mar 27;5(1):1-25. doi: doi: 10.1146/annurev.clinpsy.032408.153639.

73 de Vet HC, Terwee CB, Mokkink L, Knol DL. Responsiveness. Measurements in medicine- Practical guides to biostatistics and epidemiology. Cambridge: Cambridge University Press; 2011. p. 202-26.

74 Rieder HL, Lauritsen JM. Quality assurance of data: ensuring that numbers reflect operational definitions and contain real measurements. Int J Tuberc Lung Dis. 2011 Mar;15(3):296-304.

75 Ascher R.N., Koppelmann G.M., Miller M.J., Nagy G., Shelton G.L.Jr. An interactive system for reading unformatted printed text. IEEE Trans Comput. 1971;C-20(12):1527-43.

76 Nagy G. A preliminary investigation of techniques for the automated reading of unformatted text. Comm ACM. 1968;11(7):480-7.

77 Couch A., Keniston K. Yeasayers and naysayers: agreeing response set as a personality variable. J Abnorm Soc Psychol. 1960;60(Mar):151-74.

78 Edwards AL. The Social Desirability Variable in Personality Assessment and Research. New York: The Dryden Press; 1957.

79 Jones R.R. Differences in response consistency and subject´s preferences for three personality response formats. Proceedings of the 76th annual convention of the American Psychological Association. 1968;247-8.

80 Nishisato N., Torii Y. Effects of categorizing continous normal distributions on the product-moment correlation. Japanese Psychological Research. 1970;13:45-9.

81 Preston CC, Colman AM. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. Acta Psychol (Amst). 2000 Mar;104(1):1-15.

82 Abele L., Wahl E., Scheri W. Procedures for an automatic segmentation of text graphic and halftone regions in document. Proc.2nd Scandinavian Conf.on Image Analysis; 1981 p. 177-82.

83 Toyoda J., Noguchi Y., Nishimura Y. Study of extracting Japanese newspaper article. Proc.6th Int.Conf.on Pattern Recognition; 1982 p. 1113-5.

84 Wong I.Y., Casey R.G., Wahl E.M. Document analysis system. IBM J Reasearch Develop. 1982;26(6):647-56.

85 Kobak K.A., Mundt J.C., Greist J.H., Katzelnick D.J., Jefferson J.W. Computer assessment of depression: automating the Hamilton Depression rating scale. Drug Inf J. 2000;(34):145-56.

86 Kubick W.R. The elegant machine: applying technology to optimize the clinical trial. Drug Inf J. 1998;(32):861-9.

87 Lampe A.J., Weiler J.M. Data capture from the sponsors and investigators' perspectives: balancing quality, speed, and cost. Drug Inf J. 1988;(32):871-86.

88 Transnational Working Group on Data Management. European Clinical Research Infrastructures Network: GCP-compliant data management in multinational clinical trials. 2011.

89 Goldberg SI, Niemierko A, Turchin A. Analysis of data errors in clinical research databases. AMIA Annu Symp Proc. 2008;242-6.

90 Weir CR, Hurdle JF, Felgar MA, Hoffman JM, Roth B, Nebeker JR. Direct text entry in electronic progress notes. An evaluation of input errors. Methods Inf Med. 2003;42(1):61-7.

91 Edwards PJ, Roberts I, Clarke MJ, Diguiseppi C, Wentz R, Kwan I, et al. Methods to increase response to postal and electronic questionnaires. Cochrane Database Syst Rev. 2009;(3):MR000008.

92 Ferriss AL. A note on stimulating response to questionnaires. Am Sociol Rev. 1951;(16):247-9.

93 Streiner D.L., Norman G.R. Health Measurement Scales. Third edition ed. New York: Oxford University Press; 2003.

94 Arditi A, Cho J. Serifs and font legibility. Vision Res. 2005 Nov;45(23):2926-33.

95 Mansfield JS, Legge GE, Bane MC. Psychophysics of reading. XV: Font effects in normal and low vision. Invest Ophthalmol Vis Sci. 1996 Jul;37(8):1492-501.

96 Pettit FA. A comparison of World-Wide Web and paper-and-pencil personality questionnaires. Behav Res Methods Instrum Comput. 2002 Feb;34(1):50-4.

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

98 Murray DW, Fitzpatrick R, Rogers K, Pandit H, Beard DJ, Carr AJ, et al. The use of the Oxford hip and knee scores. J Bone Joint Surg Br. 2007 Aug;89(8):1010-4.

99 Dawson J, Fitzpatrick R, Churchman D, Verjee-Lorenz A, Clayson D. User Manual for the Oxford Hip Score (OHS) Version 1.0. Isis Innovation Limited, 2010; 2010.

100 Uesugi Y, Makimoto K, Fujita K, Nishii T, Sakai T, Sugano N. Validity and responsiveness of the Oxford hip score in a prospective study with Japanese total hip arthroplasty patients. J Orthop Sci. 2009 Jan;14(1):35-9.

101 Garbuz DS, Xu M, Sayre EC. Patients' outcome after total hip arthroplasty: a comparison between the Western Ontario and McMaster Universities index and the Oxford 12-item hip score. J Arthroplasty. 2006 Oct;21(7):998-1004.

102 Gosens T, Hoefnagels NH, de Vet RC, Dhert WJ, van Langelaan EJ, Bulstra SK, et al. The "Oxford Heup Score": the translation and validation of a questionnaire into Dutch to evaluate the results of total hip arthroplasty. Acta Orthop. 2005 Apr;76(2):204-11.

103 Kalairajah Y, Azurza K, Hulme C, Molloy S, Drabu KJ. Health outcome measures in the evaluation of total hip arthroplasties--a comparison between the Harris hip score and the Oxford hip score. J Arthroplasty. 2005 Dec;20(8):1037-41.

104 Wylde V, Learmonth ID, Cavendish VJ. The Oxford hip score: the patient's perspective. Health Qual Life Outcomes. 2005;3:66.

105 Delaunay C, Epinette JA, Dawson J, Murray D, Jolles BM. Cross-cultural adaptations of the Oxford-12 HIP score to the French speaking population. Orthopaedics & Traumatology: Surgery & Research. 2009 Apr;95(2):89-99. doi: 10.1016/j.otsr.2009.01.003.

106 Martinelli N, Longo U, Marinozzi A, Franceschetti E, Costa V, Denaro V. Cross-cultural adaptation and validation with reliability, validity, and responsiveness of the Italian version of the Oxford Hip Score in patients with hip osteoarthritis. Qual Life Res. 2011 Aug 1;20(6):923-9.

107 Naal FD, Sieverding M, Impellizzeri FM, von KF, Mannion AF, Leunig M. Reliability and validity of the cross-culturally adapted German Oxford hip score. Clin Orthop Relat Res. 2009 Apr;467(4):952-7.

108 Arden NK, Kiran A, Judge A, Biant LC, Javaid MK, Murray DW, et al. What is a good patient reported outcome after total hip replacement? Osteoarthritis Cartilage. 2011 Feb;19(2):155-62. doi: 10.1016/j.joca.2010.10.004.

109 Browne J, Jamieson L, Lewsey J, van der Meulen J, Black N, Cairns J, et al. Patient Reported Outcome Measures (PROMs) in Elective Surgery. Report to the Department of Health. 2007.

110 Nilsdotter AK, Lohmander LS, Klassbo M, Roos EM. Hip disability and osteoarthritis outcome score (HOOS)--validity and responsiveness in total hip replacement. BMC Musculoskelet Disord. 2003 May 30;4:10.

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

112 Rasch G. Probabilistic Model for Some Intelligence and Attainment Tests. Chicago: University of Chicago Press; 1960.

113 Davis AM, Perruccio AV, Canizares M, Tennant A, Hawker GA, Conaghan PG, et al. The development of a short measure of physical function for hip OA HOOS-Physical Function Shortform (HOOS-PS): an OARSI/OMERACT initiative. Osteoarthritis Cartilage. 2008 May;16(5):551-9.

114 Davis AM, Perruccio AV, Cañizares M, Hawker GA, Roos EM, Lohmander LS. Comparative Evaluation of Validity and Responsiveness of the HOOS-PS/KOOS-PS and WOMAC Following Total Joint Replacement. OARSI, Rome, Italy, September 2008.; 2008.

115 Brooks R. EuroQol: the current state of play. Health Policy. 1996 Jul;37(1):53-72.

116 The EuroQol Group. EuroQol--a new facility for the measurement of health-related quality of life. Health Policy. 1990 Dec;16(3):199-208.

117 Wittrup-Jensen KU, Lauridsen J, Gudex C, Pedersen KM. Generation of a Danish TTO value set for EQ-5D health states. Scand J Public Health. 2009 Jul;37(5):459-66.

118 Dolan P, Roberts J. Modelling valuations for Eq-5d health states: an alternative model using differences in valuations. Med Care. 2002 May;40(5):442-6.

119 Dawson J, Fitzpatrick R, Frost S, Gundle R, Lardy-Smith P, Murray D. Evidence for the validity of a patient-based instrument for assessment of outcome after revision hip replacement. J Bone Joint Surg Br. 2001 Nov;83(8):1125-9.

120 Linde L. Health-related quality of life in patients with rheumatoid arthritis A comparative validation of selected measurement instruments. Copenhagen, Denmark: Department of Rheumatology, Hvidovre Hospital, Faculty of Health Sciences, University of Copenhagen; 2009.

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

122 Gandhi SK, Salmon JW, Zhao SZ, Lambert BL, Gore PR, Conrad K. Psychometric evaluation of the 12-item short-form health survey (SF-12) in osteoarthritis and rheumatoid arthritis clinical trials. Clin Ther. 2001 Jul;23(7):1080-98.

123 Ware JEJr, Kosinski M, Bayliss MS, McHorney CA, Rogers WH, Raczek A. Comparison of methods for the scoring and statistical analysis of SF-36 health profile and summary measures: summary of results from the Medical Outcomes Study. Med Care. 1995 Apr;33(4 Suppl):AS264-AS279.

124 Singh J, Sloan JA, Johanson NA. Challenges With Health-related Quality of Life Assessment in Arthroplasty Patients: Problems and Solutions. J Am Acad Orthop Surg. 2010 Feb 1;18(2):72-82.

125 Ahmad MA, Xypnitos FN, Giannoudis PV. Measuring hip outcomes: Common scales and checklists. Injury. 2011 Mar;42(3):259-64.

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

127 World Health Organization. Process of translation and adaptation of instruments. 2010 [cited 2011 Nov 5]. [Internet]. World Health Organization. Available from: http://www.who.int/substance_abuse/research_tools/translation/en/.

128 Quintana JM, Aguirre U, Barrio I, Orive M, Garcia S, Escobar A. Outcomes after total hip replacement based on patients' baseline status: What results can be expected? Arthritis Care Res. 2012;64(4):563-72. doi: 10.1002/acr.21570.

129 Tubach F, Giraudeau B, Ravaud P. The variability in minimal clinically important difference and patient acceptable symptomatic state values did not have an impact on treatment effect estimates. J Clin Epidemiol. 2009 Jul 1;62(7):725-8.

130 Maksymowych WP, Gooch K, Dougados M, Wong RL, Chen N, Kupper H, et al. Thresholds of patient-reported outcomes that define the patient acceptable symptom state in ankylosing spondylitis vary over time and by treatment and patient characteristics. Arthritis Care Res (Hoboken). 2010 Jun;62(6):826-34.

131 Keurentjes JC, Van Tol FR, Fiocco M, Schoones JW, Nelissen RG. Minimal clinically important differences in health-related quality of life after total hip or knee replacement. Bone and Joint Research. 2012 May 1;1(5):71-7.

132 King MT. A point of minimal important difference (MID): a critique of terminology and methods. Expert Rev Pharmacoeconomics Outcomes Res. 2011 Apr 1;11(2):171-84. doi: 10.1586/erp.11.9.

133 Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol. 2008 Feb;61(2):102-9. doi: 10.1016/j.jclinepi.2007.03.012.

134 Shi HY, Chang JK, Wong CY, Wang JW, Tu YK, Chiu HC, et al. Responsiveness and minimal important differences after revision total hip arthroplasty. BMC Musculoskeletal Disorders. 2010;11(1):261. doi: 10.1186/1471-2474-11-261.

135 Dougados M, Brault Y, Logeart I, van der Heijde D, Gossec L, Kvien T. Defining cut-off values for disease activity states and improvement scores for patient-reported outcomes: the example of the Rheumatoid Arthritis Impact of Disease (RAID). Arthritis Research & Therapy. 2012;14(3):R129. doi: 10.1186/ar3859.

136 Davis AM, Perruccio AV, Canizares M, Tennant A, Hawker GA, Conaghan PG, et al. The development of a short measure of physical function for hip OA HOOS-Physical Function Shortform (HOOS-PS): an OARSI/OMERACT initiative. Osteoarthritis Cartilage. 2008 May;16(5):551-9.

137 Heiberg T, Kvien TK, Mowinckel P, Aletaha D, Smolen JS, Hagen KB. Identification of disease activity and health status cut-off points for the symptom state acceptable to patients with rheumatoid arthritis. Ann Rheum Dis. 2008 Jul 1;67(7):967-71.

138 Kvamme MK, Kristiansen IS, Lie E, Kvien TK. Identification of Cutpoints for Acceptable Health Status and Important Improvement in Patient-Reported Outcomes, in Rheumatoid Arthritis, Psoriatic Arthritis, and Ankylosing Spondylitis. The Journal of Rheumatology. 2010 Jan 1;37(1):26-31.

139 Kvien TK, Heiberg T, Hagen KB. Minimal clinically important improvement/difference (MCII/MCID) and patient acceptable symptom state (PASS): what do these concepts mean? Ann Rheum Dis. 2007 Nov;66 Suppl 3:iii40-iii41.

140 Maksymowych WP, Richardson R, Mallon C, van der Heijde D+, Boonen A. Evaluation and validation of the patient acceptable symptom state (PASS) in patients with ankylosing spondylitis. Arthritis Care Res. 2007;57(1):133-9. doi: 10.1002/art.22469.

141 Froud RJ. ROCMIC: Stata module to estimate minimally important change (MIC) thresholds for continuous clinical outcome measures using ROC curves [computer program]. 2002.

142 Norman GR, Sloan JA, Wyrwich KW. Interpretation of Changes in Health-Related Quality of Life: The Remarkable Universality of Half a Standard Deviation. Med Care. 2003 May 1;41(5):582-92.

143 Myles PS, Cui J. I. Using the Bland-Altman method to measure agreement with repeated measures. Br J Anaesth. 2007 Sep 1;99(3):309-11.

144 Beaton DE, Bombardier C, Katz JN, Wright JG, Wells G, Boers M, et al. Looking for important change/differences in studies of responsiveness. OMERACT MCID Working Group. Outcome Measures in Rheumatology. Minimal Clinically Important Difference. The Journal of Rheumatology. 2001 Feb 1;28(2):400-5.

145 de Vet HC, Terwee CB, Ostelo RW, Beckerman H, Knol DL, Bouter LM. Minimal changes in health status questionnaires: Distinction between minimally detectable change and minimally important change. Health Qual Life Outcomes. 2006;4:54. doi: 10.1186/1477-7525-4-54.

146 Gartner FR, Nieuwenhuijsen K, van Dijk FJ, Sluiter JK. Interpretability of change in the Nurses Work Functioning Questionnaire: minimal important change and smallest detectable change. J Clin Epidemiol. 2012 Dec;65(12):1337-47.

147 Cella D, Eton DT, Lai JS, Peterman AH, Merkel DE. Combining Anchor and Distribution-Based Methods to Derive Minimal Clinically Important Differences on the Functional Assessment of Cancer Therapy (FACT) Anemia and Fatigue Scales. Journal of Pain and Symptom Management 24[6], 547-561. 1-12-2002.

148 Cohen J. Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.; 1988.

149 Kazis LE, Anderson JJ, Meenan RF. Effect Sizes for Interpreting Changes in Health Status. Med Care. 1989 Mar 1;27(3):S178-S189.

150 Walters SJ, Brazier JE. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. Qual Life Res. 2005 Aug;14(6):1523-32.

151 Bond M, Davis A, Lohmander S, Hawker G. Responsiveness of the OARSI-OMERACT osteoarthritis pain and function measures. Osteoarthritis Cartilage. 2012 Jun;20(6):541-7.

152 Stucki G, Liang MH, Fossel AH, Katz JN. Relative responsiveness of condition-specific and generic health status measures in degenerative lumbar spinal stenosis. J Clin Epidemiol. 1995 Nov;48(11):1369-78.

153 Liang MH, Fossel AH, Larson MG. Comparisons of Five Health Status Instruments for Orthopedic Evaluation. Med Care. 1990 Jul 1;28(7):632-42.

154 Hays RD, Farivar SS, Liu H. Approaches and Recommendations for Estimating Minimally Important Differences for Health-Related Quality of Life Measures. COPD. 2005 Jan 1;2(1):63-7. doi: 10.1081/COPD-200050663.

155 Revicki DA, Erickson PA, Sloan JA, Dueck A, Guess H, Santanello NC, et al. Interpreting and Reporting Results Based on Patient-Reported Outcomes. Value in Health. 2007;10:S116-S124. doi: 10.1111/j.1524-4733.2007.00274.x.

156 Terwee CB, Roorda LD, Dekker J, Bierma-Zeinstra SM, Peat G, Jordan KP, et al. Mind the MIC: large variation among populations and methods. J Clin Epidemiol. 2010 May;63(5):524-34.

157 Wyrwich KW, Tierney WM, Wolinsky FD. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. J Clin Epidemiol. 1999 Sep;52(9):861-73.

158 Yost KJ, Eton DT. Combining Distribution- and Anchor-Based Approaches to Determine Minimally Important Differences. Evaluation & the Health Professions. 2005 Jun 1;28(2):172-91.

159 Klassbo M, Larsson E, Mannevik E. Hip disability and osteoarthritis outcome score An extension of the Western Ontario and McMaster Universities Osteoarthritis Index. Scand J Rheumatol. 2003 Jan 1;32(1):46-51. doi: 10.1080/03009740310000409.

160 Ornetti P, Perruccio AV, Roos EM, Lohmander LS, Davis AM, Maillefert JF. Psychometric properties of the French translation of the reduced KOOS and HOOS (KOOS-PS and HOOS-PS). Osteoarthritis Cartilage. 2009 Dec;17(12):1604-8. doi: 10.1016/j.joca.2009.06.007.

161 Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol. 2003 May 1;56(5):395-407.

162 de Vet HCW, Terluin B, Knol DL, Roorda LD, Mokkink LB, Ostelo RWJG, et al. Three ways to quantify uncertainty in individually applied "minimally important change" values. J Clin Epidemiol. 2010 Jan;63(1):37-45.

163 Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. J Consult Clin Psychol. 1991 Feb;59(1):12-9.

164 Guyatt GH, Bombardier C, Tugwell PX. Measuring disease-specific quality of life in clinical trials. CMAJ. 1986 Apr 15;134(8):889-95.

165 Liang MH, Larson MG, Cullen KE, Schwartz JA. Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. Arthritis Rheum. 1985 May;28(5):542-7.

166 Tubach F, Ravaud P, Baron G, Falissard B, Logeart I, Bellamy N, et al. Evaluation of clinically relevant changes in patient reported outcomes in knee and hip osteoarthritis: the minimal clinically important improvement. Ann Rheum Dis. 2005 Jan 1;64(1):29-33.

167 Kamper-Jørgensen F. Danskernes sundhed 2005 Skema 2. Statens Institut for Folkesundhed, Socialforskningsinstituttet; 2005.

168 Clinical Effectiveness Unit. Questionnaire for patients who have had hip surgery. POIS Audit, The Royal College of Surgeons of England; 2009.

169 Chard J, Kuczawski M, van der Meulen J. Patient Outcomes in Surgery. A report comparing Independent Sector Treatment Centres and NHS providers. POiS Audit Steering Committee, Clinical Effectiveness Unit, The Royal College of Surgeons of England; 2011.

170 Wineberg A. Finalised Patient Reported Outcome Measures (PROMs) in England. April 2010 to March 2011. Pre- and post-operative data. SUS/HES Analysis (Development) team at the Health and Social Care Information Centre; 2012.

171 Augestad LA, Rand-Hendriksen K. Influence of construct-irrelevant factors and effects of methodological choices on EQ-5D health state valuation. Oslo, Norway: Faculty of Medicine, University of Oslo; 2012.

172 Cheung Kajang, Oemar Mandy, Oppe Mark, Rabin Rosalin. EQ-5D User Guide. Basic information on how to use EQ-5D. Version 2.0. March 2009. The EuroQoL Group; 2009.

173 Ware JE, Kosinski M, Turner-Bowker DM, Gandek B. User`s Manual for the SF-12v2 Health Survey (With a supplement Documenting the SF-12 Health Survey). Lincoln, Rhode Island: QualityMetric Incorporated; 2009.

174 Roos EM. A User's Guide to: Hip disability and Osteoarthritis Outcome Score HOOS. Updated May 2008. 2003.

175 Mitchell MN. wtest [computer program]. Statistical Computing and Consulting, UCLA, Academic Technology Services; 2000.

176 Wilcox RR, in VL, Thompson KL. New monte carlo results on the robustness of the anova f, w and f statistics. Communications in Statistics - Simulation and Computation. 1986 Jan 1;15(4):933-43. doi: doi: 10.1080/03610918608812553.

177 de Vet HC, Terwee CB, Mokkink L, Knol DL. Reliability. Measurements in medicine- Practical guides to biostatistics and epidemiology. Cambridge: Cambridge University Press; 2011. p. 96-149.

178 Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979 Mar;86(2):420-8.

179 de Vet HC, Terwee CB, Mokkink L, Knol DL. Interpretability. Measurements in medicine- Practical guides to biostatistics and epidemiology.Cambridge: Cambridge University Press; 2011. p. 227-68.

180 Davis AM, Perruccio AV, Canizares M, Tennant A, Hawker GA, Conaghan PG, et al. The development of a short measure of physical function for hip OA HOOS-Physical Function Shortform (HOOS-PS): an OARSI/OMERACT initiative. Osteoarthritis Cartilage. 2008 May;16(5):551-9.

181 Tubach F, Dougados M, Falissard B, Baron G, Logeart I, Ravaud P. Feeling good rather than feeling better matters more to patients. Arthritis Care Res. 2006;55(4):526-30. doi: 10.1002/art.22110.

182 Campbell MK, Torgerson DJ. Bootstrapping: estimating confidence intervals for cost-effectiveness ratios. QJM. 1999 Mar;92(3):177-82.

183 Fiellin DA, Feinstein AR. Bootstraps and jackknives: new, computer-intensive statistical tools that require no mathematical theories. J Investig Med. 1998 Feb;46(2):22-6.

184 de Vet HC, Terwee CB, Mokkink L, Knol DL. Field-testing: item reduction and data structure. Measurements in medicine- Practical guides to biostatistics and epidemiology.Cambridge: Cambridge University Press; 2011. p. 65-95.

185 de Groot IB, Reijman M, Terwee CB, Bierma-Zeinstra SMA, Favejee M, Roos EM, et al. Validation of the Dutch version of the Hip disability and Osteoarthritis Outcome Score. Osteoarthritis Cartilage. 2007 Jan;15(1):104-9. doi: 10.1016/j.joca.2006.06.014.

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

187 Dawson J, Fitzpatrick R, Frost S, Gundle R, Lardy-Smith P, Murray D. Evidence for the validity of a patient-based instrument for assessment of outcome after revision hip replacement. J Bone Joint Surg Br. 2001 Nov;83(8):1125-9.

188 Dawson J, Fitzpatrick R, Frost S, Gundle R, Lardy-Smith P, Murray D. Evidence for the validity of a patient-based instrument for assessment of outcome after revision hip replacement. J Bone Joint Surg Br. 2001 Nov;83(8):1125-9.

189 Wylde V, Blom AW, Whitehouse SL, Taylor AH, Pattison GT, Bannister GC. Patient-Reported Outcomes After Total Hip and Knee Arthroplasty: Comparison of Midterm Results. The Journal of Arthroplasty. 2009 Feb;24(2):210-6. doi: 10.1016/j.arth.2007.12.001.

190 Busija L, Osborne R, Nilsdotter A, Buchbinder R, Roos E. Magnitude and meaningfulness of change in SF-36 scores in four types of orthopedic surgery. Health and Quality of Life Outcomes. 2008;6(1):55. doi: 10.1186/1477-7525-6-55.

191 Ostendorf M, van Stel HF, Buskens E, Schrijvers AJ, Marting LN, Verbout AJ, et al. Patient-reported outcome in total hip replacement. A comparison of five instruments of health status. J Bone Joint Surg Br. 2004 Aug;86(6):801-8.

192 Linde L. Health-related quality of life in patients with rheumatoid arthritis A comparative validation of selected measurement instruments. Copenhagen, Denmark: Department of Rheumatology, Hvidovre Hospital, Faculty of Health Sciences, University of Copenhagen; 2009.

193 Davis AM, Perruccio AV, Lohmander LS. Minimally clinically important improvement: all non-responders are not really non-responders an illustration from total knee replacement. Osteoarthritis Cartilage. 2012 May;20(5):364-7.

194 de Vet HC, Terwee CB, Mokkink L, Knol DL. Interpretability. Measurements in medicine- Practical guides to biostatistics and epidemiology.Cambridge: Cambridge University Press; 2011. p. 227-68.

195 Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393.

196 Nilsdotter AK Lohmander LS, Klassbo M, Roos EM. Hip disability and osteoarthritis outcome score (HOOS)–validity and responsiveness in total hip replacement. BMC Musculoskelet Disord 2003; 4: 10.

197 Jorgensen C, Karlsmose B. Validation of automated forms processing A comparison of TeleformTM with manual data entry. Comput Biol Med. 1998 Nov;28(6):659-67. doi: 10.1016/S0010-4825(98)00038-9.

198 Weller SC, Baer RD. Using Electronic Scanning Forms for Data Entry. Field Methods. 2001 May 1;13(2):198-203.

199 Cooke DJ, Michie C, Hart SD, Hare RD. Evaluating the Screening Version of the Hare Psychopathy Checklist-Revised (PCL:SV): An Item Response Theory Analysis. [Article]. Psychological Assessment March 1999;11(1):3-13.(1):3-13.

200 Davis AM. The development of a short measure of physical function for hip OA HOOS-Physical Function Shortform (HOOS-PS): an OARSI/OMERACT initiative. 2008 May.

201 Nilsdotter AK Lohmander LS, Klassbo M, Roos EM. Hip disability and osteoarthritis outcome score (HOOS)–validity and responsiveness in total hip replacement. BMC Musculoskelet Disord 2003; 4: 10.

202 Wolinsky FD, Wyrwich KW, Nienaber NA, Tierney WM. Generic versus Disease-Specific Health Status Measures: An Example Using Coronary Artery Disease and Congestive Heart Failure Patients. Evaluation & the Health Professions. 1998 Jun 1;21(2):216-43.

203 Choi SW, Gibbons LE, Crane PK. lordif: An R Package for Detecting Differential Item Functioning Using Iterative Hybrid Ordinal Logistic Regression/Item Response Theory and Monte Carlo Simulations. J Stat Softw. 2011 Mar 1;39(8):1-30.

204 Crane PK, Gibbons LE, Jolley L, van BG. Differential item functioning analysis with ordinal logistic regression techniques. DIFdetect and difwithpar. Med Care. 2006 Nov;44(11 Suppl 3):S115-S123.

205 Zumbo BD. A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert type (ordinal) item scores. Ottawa, ON: 1999.

206 Pollard B, Johnston M, Dixon D. Exploring differential item functioning in the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC). BMC Musculoskeletal Disorders. 2012;13(1):265. doi: 10.1186/1471-2474-13-265.

207 Singh JA, Lewallen D. Age, gender, obesity, and depression are associated with patient-related pain and function outcome after revision total hip arthroplasty. Clin Rheumatol. 2009;28(12):1419-30. Pubmed PMID: PMC2963019.

208 Jones CA, Voaklander DC, Johnston DW, Suarez-Almazor ME. The Effect of Age on Pain, Function, and Quality of Life After Total Hip and Knee Arthroplasty. Arch Intern Med. 2001 Feb 12;161(3):454-60.

209 Santaguida PL, Hawker GA, Hudak PL, Glazier R, Mahomed NN, Kreder HJ, et al. Patient characteristics affecting the prognosis of total hip and knee joint arthroplasty: a systematic review. Can J Surg. 2008;51(6):428-36. Pubmed PMID: PMC2592576.

210 Lee YK, Chung C, Park M, Lee K, Lee D, Lee S, et al. Transcultural adaptation and testing of psychometric properties of the Korean version of the Oxford hip score. J Orthop Sci. 2012;17(4):377-81.

211 Tubach F, Ravaud P, Martin-Mola E, Awada H, Bellamy N, Bombardier C, et al. Minimum clinically important improvement and patient acceptable symptom state in pain and function in rheumatoid arthritis, ankylosing spondylitis, chronic back pain, hand osteoarthritis, and hip and knee osteoarthritis: Results from a prospective multinational study. Arthritis Care Res. 2012;64(11):1699-707. doi: 10.1002/acr.21747.

212 Escobar A, Gonzalez M, Quintana JM, Vrotsou K, Bilbao A, Herrera-Espineira C, et al. Patient acceptable symptom state and OMERACT - OARSI set of responder criteria in joint replacement. Identification of cut-off values. Osteoarthritis and cartilage / OARS, Osteoarthritis Research Society. 2012 Feb 1;20(2):87-92.

213 Browne JP, van der Meulen JH, Lewsey JD, Lamping DL, Black N. Mathematical coupling may account for the association between baseline severity and minimally important difference values. J Clin Epidemiol. 2010 Aug;63(8):865-74. doi: 10.1016/j.jclinepi.2009.10.004.

214 Santanello NC, Zhang J, Seidenberg B, Reiss TF, Barber BL. What are minimal important changes for asthma measures in a clinical trial? Eur Respir J. 1999 Jul 1;14(1):23-7.

215 Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR. Methods to Explain the Clinical Significance of Health Status Measures. Mayo Clin Proc. 2002 Apr;77(4):371-83.

216 Barber BL, Santanello NC, Epstein RS. Impact of the Global on Patient Perceivable Change in an Asthma Specific QOL Questionnaire. Qual Life Res. 1996 Feb 1;5(1):117-22.

217 Guyatt GH, Norman GR, Juniper EF, Griffith LE. A critical look at transition ratings. J Clin Epidemiol. 2002 Sep;55(9):900-8.

218 Hurst NP, 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). Rheumatology (Oxford). 1997 May 1;36(5):551-9.

219 Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters. 2006 Jun;27(8):861-74. doi: doi: 10.1016/j.patrec.2005.10.010.

220 Beaton DEBP. Simple as Possible? Or Too Simple?: Possible Limits to the Universality of the One Half Standard Deviation. [Miscellaneous]. Med Care. 2003 May;41(5):593-6.

221 Farivar SS, Liu H, Hays RD. Half standard deviation estimate of the minimally important difference in HRQOL scores? Expert Rev Pharmacoeconomics Outcomes Res. 2004 Oct 1;4(5):515-23. doi: doi: 10.1586/14737167.4.5.515.

222 Wright JGM. Interpreting Health-Related Quality of Life Scores: The Simple Rule of Seven May not be so Simple. [Miscellaneous]. Med Care. 2003 May;41(5):597-8.

223 Nilsdotter AK Lohmander LS, Klassbo M, Roos EM. Hip disability and osteoarthritis outcome score (HOOS)–validity and responsiveness in total hip replacement. BMC Musculoskelet Disord 2003; 4: 10.

224 Norquist JM, Fitzpatrick R, Jenkinson C. Health-related quality of life in amyotrophic lateral sclerosis: determining a meaningful deterioration. Qual Life Res. 2004 Oct;13(8):1409-14.

225 Crosby RD, Kolotkin RL, Williams GR. An integrated method to determine meaningful changes in health-related quality of life. J Clin Epidemiol. 2004 Nov;57(11):1153-60.

226 Ferguson RJ, Robinson AB, Splaine M. Use of the reliable change index to evaluate clinical significance in SF-36 outcomes. Qual Life Res. 2002 Sep;11(6):509-16.

227 Giesinger JM, Kuster MS, Behrend H, Giesinger K. Association of psychological status and patient-reported physical outcome measures in joint arthroplasty: a lack of divergent validity. Health Qual Life Outcomes. 2013;11:64.

228 Hoeymans N, Feskens EJM, van den Bos GAM, Kromhout D. Measuring functional status: Cross-sectional and longitudinal associations between performance and self-report (Zutphen Elderly Study 1990GÇô1993). J Clin Epidemiol. 1996 Oct;49(10):1103-10.

229 Jansson KA, Granath F. Health-related quality of life (EQ-5D) before and after orthopedic surgery. Acta Orthopaedica. 2010 Dec 29;82(1):82-9. doi: 10.3109/17453674.2010.548026.

230 KRAM-enheden. KRAM-spørgeskema (papir version). Statens Institut for Folkesundhed; 2007.

231 Lauritsen J. Simpel Funktionsmåling - instruksjonsmateriale. DVD og pjece. Odense Universitets Hospital; 2007.

232 Burr H, Olsen O. Helbred, livsstil og arbejdsmiljø Spørgeskema 2005. AMI; 2005.

233 Region Nordjylland. Spørgeskema - til helbredssamtale ved natarbejde. Koncern HR-Forhandling og Arbejdsmiljø; 2010.

234 Sandsjo L, Larsman P, Vollenbroek-Hutten M, Laubli T, Juul-Kristensen B, Klipstein A, et al. Comparative assessment of study groups of elderly female computer users from four European countries: questionnaires used in the NEW study. European Journal of Applied Physiology. 2006 Jan 1;96(2):122-6.