MEASURING PATIENT-REPORTED PHYSICAL FUNCTION AND PAIN IN TOTAL **HIP** AND **KNEE** ARTHROPLASTY

CHRISTEL BRAAKSMA

Measuring patient-reported physical function and pain in total hip and knee arthroplasty

Christel Braaksma

The studies presented in this thesis were conducted at the Department of Orthopedic surgery, St. Antonius Hospital, Utrecht. This research in this thesis was embedded in Amsterdam Movement Sciences Research Institute, at the Department of Health Sciences, Vrije Universiteit Amsterdam.

The printing of this thesis was financially supported by the Maatschap Orthopedie St. Antonius Ziekenhuis, Dutch Orthopedic Society (NOV), St. Antonius Hospital, Artrose Instituut Nederland, Link Lima Nederland, Chipsoft and Amsterdam Movement Sciences Research Institute.

Cover and chapters:	conceptual design and art - Jeannette Nijhuis, graphic design -
Lay-out:	Tiny Wouters
Printed by:	Proefschriftspecialist

https://doi.org/10.5463/thesis.1659

VRIJE UNIVERSITEIT

MEASURING PATIENT-REPORTED PHYSICAL FUNCTION AND PAIN IN TOTAL HIP AND KNEE ARTHROPLASTY

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. J.J.G. Geurts,
volgens besluit van de decaan
van de Faculteit der Bètawetenschappen
in het openbaar te verdedigen
op vrijdag 12 juni 2026 om 11.45 uur
in de universiteit

door

Christel Braaksma

geboren te Noordoostpolder

Promotoren	prof. dr. R.W.J.G. Ostelo prof. dr. C.B. Terwee
Copromotoren	dr. N. Wolterbeek dr. M.R. Veen
Beoordelingscommissie	prof. dr. J.E. Bosmans dr. M.A.H. Oude Voshaar dr. D.O. Verbeek MBA prof. dr. P.J. van der Wees prof. dr. T.P.M. Vliet Vlieland prof. dr. T. Gosens

Table of contents

CHAPTER 1

General introduction,
aim and outline of this thesis

Total hip and total knee arthroplasty

Total hip arthroplasty (THA) and total knee arthroplasty (TKA) are surgical procedures primarily aimed at improving function and reducing pain for patients with end-stage osteoarthritis who have not responded adequately to non-operative management[1,2]. In the Netherlands, the volume of THA and TKA procedures has increased substantially over the past few decades. In 2024, approximately 50,000 THA and 40,000 TKA procedures were performed[3]. Currently, about one in twelve Dutch inhabitants has undergone at least one hip, knee, or shoulder replacement[4].

Both THA and TKA are commonly performed and highly successful interventions. The primary indication for these surgeries is symptomatic, advanced osteoarthritis associated with persistent pain and substantial limitations in physical functioning despite adequate conservative treatment[5]. The primary goals of THA and TKA are to achieve lasting pain relief, restore function and mobility, and improve health-related quality of life.

Success in these procedures is commonly assessed using a combination of factors: objective implant survival (time to revision), surgeon-reported measures (such as range of motion, X-rays for the hardware alignment, and length of hospital stay), and patient-reported outcome measures (PROMs), e.g., for self-reported pain and physical functioning[6]. Accurately assessing experienced pain and physical function is essential for evaluating treatment effectiveness in patients undergoing arthroplasty, since these domains are the primary indicators of THA and TKA[5].

PROMs are designed to capture patients' perspectives on these health domains[7]. Postoperative changes in PROM scores related to pain and physical function are therefore considered the most relevant measures of treatment success, since PROMs uniquely reflect what matters most to patients[6,7]. Consequently, the implementation of reliable and valid PROMs in research and clinical practice is essential, as both patients and healthcare providers rely on accurate evaluations.

In the Netherlands, key characteristics related to THA and TKA procedures are systematically collected and maintained within the Dutch Arthroplasty Register (Landelijke Registratie Orthopedische Interventies; LROI). The LROI is a national clinical registry established in 2007 to monitor the volume, practice variation, and outcomes of arthroplasty procedures nationwide[3,8]. Launched by the Dutch Orthopedic Society (NOV), a comprehensive nationwide collection of PROM data began in 2013 within the

LROI. The administration of these PROMs is now a mandatory quality indicator for every orthopedic care clinic.

The NOV recommends a standardized set of PROMs for patients undergoing THA and TKA, administered at multiple time points before and after surgery[9]. This set includes generic PROMs (numeric rating scales measuring pain and the EQ-5D-5L[10]) and disease-specific PROMs. At the start of this project, the Hip disability and Osteoarthritis Outcome Score- Physical function Shortform (HOOS-PS; [11]) and the Knee disability and Osteoarthritis Outcome Score- Physical function Shortform (KOOS-PS; [12]) were mandatory disease-specific PROMs. These instruments aim to measure physical function in patients with hip or knee osteoarthritis, respectively.

Possible advantages of PROMs

The use of PROMs offers numerous potential benefits for patients, clinical practice, research, and policy-making.

Primarily, for *patients,* PROMs can enhance engagement in care, support shared decision-making, strengthen the patient-provider relationship, and improve satisfaction by managing expectations[13–16]. PROM scores may also be predictive; for instance, patients reporting severe pain and functional limitations, are more likely to benefit from orthopedic intervention[17]. Platforms such as ”Patients Like Me” illustrate how PROM data can empower patients by providing personalized prognoses for orthopedic surgery[18].

Second, within *clinical settings*, PROMs support decision-making by providing healthcare professionals with a more comprehensive understanding of patients’ health status. This facilitates improved shared decision-making and quality of care[13]. Moreover, PROMs may assist in determining the optimal timing for surgery, aiming to perform arthroplasties in the “Goldilocks Zone”, when daily life is truly affected while overall health is still adequate[19]. Additionally, PROMs aid in managing postoperative expectations[14]. Other applications of PROMs in clinical settings include monitoring health outcomes and capturing patients’ perspectives on symptoms and function at multiple time points. This data can be used to track outcomes over time (e.g., pre- and post-operatively). Furthermore, PROMs can help standardize care protocols, allowing surgeons to reduce unwarranted care variation[14]. Other clinical applications include monitoring red flags (e.g. detection of complications by early symptom change using

PROMs, which might prompt clinical follow-up), and providing metrics for tailoring treatments at the individual patient level[13,14]. PROMs may also facilitate hybrid care models, such as replacing routine outpatient visits with video consultations. For example, if the reported pain or functional scores fall below expected recovery thresholds, the system flags the patient for follow-up via telehealth rather than requiring an in-person visit. PROMs can also be used as a triage tool to allocate THA patients to hospital or video consultation 6 weeks postoperatively, which can possible reduce 70% of the outpatients visits[20]. Furthermore, PROMs may also facilitate establishing cut-off values for risk assessment, clinical decision-making, and referral decisions to the orthopedic department[21].

Third, the advantage of PROMs for *academic interest* is especially for knowledge gain. They can be used for outcome predictions, treatment comparisons, and assessing the relative effectiveness and efficacy of interventions, while optimally taking the patient perspective into account.

Lastly, for *policy-making* at the levels of healthcare institutions and government agencies, PROMs support quality assurance, benchmarking, public reporting, and healthcare transparency[22].

There is a growing trend toward PROM-based performance measurement in orthopedics, complementing traditional process measures such as infection rates or readmissions[16,23]. As healthcare systems transition toward patient-centered and evidence-based care, PROMs provide quantifiable, patient-relevant outcome data. This aids diagnosis and treatment evaluation, insights inform baseline function, evaluate the effect of interventions, and may reduce unwarranted variations in care.

PROMs also play a central role in Value-Based Healthcare (VBHC), which defines value as the health outcomes achieved per euro spent, from the patient's perspective[24]. Within this framework, PROMs operationalize patient-relevant outcomes and shift the focus from the volume of services provided to the value of care delivered that truly matters to patients. Emphasizing value over volume may improve quality and efficiency while potentially addressing two major systemic challenges: healthcare's environmental impact and the healthcare workforce shortage.

VBHC has demonstrated potential to reduce healthcare costs for degenerative knee disease and arthroplasty by identifying and eliminating low-value care (e.g., unnecessary tests, procedures, and hospitalizations)[25,26]. This is particularly relevant given the

anticipated shortage of healthcare workers, in which the demand is expected to outpace supply due to an aging population and the growing prevalence of osteoarthritis[27–29]. This can affect the quality, accessibility, and affordability of care[29]. Long-term resource constraints require a shift in how resource allocation decisions are made. VBHC in orthopedic care may help determine the highest value possible in healthcare delivery during times of shortage. Furthermore, as stated above, PROM-based remote monitoring and telehealth have the potential to reduce the need for in-person visits, thereby minimizing patient travel and resource use[16]. Virtual care and the reduction of low-value care can contribute to more sustainable orthopedic care delivery.

Current shortcomings in PROMs

Despite their advantages, the measurement properties of PROMs remain a subject of debate. The currently recommended PROMs by the NOV ('legacy PROMs') have substantial concerns regarding their measurement properties. The most substantial problem with the HOOS-PS and KOOS-PS is their limited content validity, meaning they do not sufficiently capture the intended construct and include irrelevant items for patients[30,31]. In response, the NOV revised its disease-specific PROM recommendation in 2020, favoring the Oxford Hip Score (OHS; [32]) and the Oxford Knee Score (OKS; [33]). However, these PROMs also have limitations, including a measurement error that is often too large for individual assessment, insufficient responsiveness of the OKS, licensing requirements, and a fixed patient burden of 12 items per PROM[34,35]. Additionally, the content validity of the OHS is considered insufficient[36]. More broadly, legacy PROMs often have limitations in measuring the full spectrum of a health domain, resulting in floor and ceiling effects. This raises concerns about their utility, reliability, and validity in diverse clinical settings. Consequently, no currently used PROM meets all the criteria for valid and reliable outcome assessment in THA and TKA patients.

Beyond measurement limitations, PROM implementation in routine orthopedic practice is often poor and inconsistent. Studies document wide variability and low response rates[37,38]. PROMs have traditionally been viewed primarily as research tools rather than components of clinical care and quality indicators, contributing to slow adoption and poor integration in clinical practice[39]. The failure to integrate PROMs into clinical workflows underscores that, without coordinated investment, they risk remaining underutilized in orthopedics, despite their potential to improve patient-centered, value-based care.

Current concepts and perspectives in PROMs

To improve the use of PROMs, modern measurement strategies are advocated. The adoption of modern measurement strategies based on advanced psychometric methods, particularly Item Response Theory (IRT) models, is proposed. IRT enables precise, adaptive, and linear measurement by calibrating a wide range of items within a single health domain along a measurement interval scale, ordered by their difficulty[40,41]. Computer Adaptive Testing (CAT), in which the items that are presented to patients are dynamically tailored for each respondent by selecting items that provide the most relevant information based on previous responses[42]. As a result, CAT leads to more tailored PROMs, reducing the number of items to those most relevant for the patient. IRT-based assessments are concise, adaptable, efficient, reduce response burden, sustainable, and preserve measurement precision on a standardized scoring framework. Also, the Dutch Orthopedic Association has issued the 'PROMs advice 2020' that contains a proposal for suggested improvements in this area[9]. One key opportunity for improving the use of PROMs is investigating the feasibility of using CAT.

The most extensively validated instrument that combines IRT and CAT is the Patient-Reported Outcomes Measurement Information System (PROMIS®)[43]. Developed with large funding from the National Institutes of Health (NIH), PROMIS contains instruments for evaluating patient-reported health status across physical, mental, and social health domains. It offers both fixed short forms (SF) and CAT administration, providing T-score-based scales that facilitate comparability across different conditions and populations. PROMIS is proposed as an alternative to legacy PROMs because it addresses several common limitations. First, PROMIS typically requires fewer questions, especially when administered as a CAT (an average of 5 to 6 items), and less time to complete (1 to 2minutes compared to several minutes) than legacy PROMs, while maintaining comparable reliability[44]. This reduces both respondent burden and administrative workload and enhances the feasibility of routine use. The adoption of PROMIS also has the potential to address the lack of validity of legacy PROMs in THA and TKA patients, as PROMIS has demonstrated robust content validity, structural validity, a broad measurement range, responsiveness, and measurement precision[45–50]. The item banks cover the full spectrum of the health domain, minimizing the likelihood of floor and ceiling effects. Furthermore, PROMIS is a disease-transcending generic measurement instrument, thereby improving comparability across diseases and nations, and helping eliminate instrument heterogeneity[51]. As a result, PROMIS is increasingly being implemented in orthopedic practice[52]. To improve the use of PROMs, modern measurement strategies are advocated. The adoption of modern measurement

strategies based on advanced psychometric methods, particularly Item Response Theory (IRT) models, is proposed. IRT enables precise, adaptive, and linear measurement by calibrating a wide range of items within a single health domain along a measurement interval scale, ordered by their difficulty[40,41].

Computer Adaptive Testing (CAT), in which the items that are presented to patients are dynamically tailored for each respondent by selecting items that provide the most relevant information based on previous responses[42]. As a result, CAT leads to more tailored PROMs, reducing the number of items to those most relevant for the patient. IRT-based assessments are concise, adaptable, efficient, reduce response burden, sustainable, and preserve measurement precision on a standardized scoring framework. Also, the Dutch Orthopedic Association has issued the 'PROMs advice 2020' that contains a proposal for suggested improvements in this area[9]. One key opportunity for improving the use of PROMs is investigating the feasibility of using CAT.

Research gaps

Despite these advances, relatively few studies have directly compared the psychometric properties of PROMIS and legacy PROMs in THA and TKA patients[47,53]. Evidence regarding floor and ceiling effects, smallest detectable change, and responsiveness of PROMIS in orthopedic populations remains limited[54]. Furthermore, the theoretical advantages of PROMIS CAT and short forms, providing less burdensome and more relevant questionnaires, need to be validated in clinical settings, including cross-cultural validity in non-U.S. populations.

Furthermore, if a transition from legacy PROMs towards PROMIS is to occur, robust crosswalks are required to enable score conversions and longitudinal comparisons. Existing crosswalks can be utilized to facilitate score conversion by mapping legacy PROM scores to PROMIS scores. This process aids in analyzing group-level data, ensures a smoother transition, and preserves data integrity while allowing meaningful comparisons[55]. While crosswalks have previously been developed in the U.S. population between the HOOS-PS,KOOS-PS, towards PROMIS Physical Function short form scores[56–59], these have not been externally validated across different countries, health conditions, or languages. Furthermore, no crosswalks currently exist to transform scores from the currently used Oxford Hip Score and Oxford Knee Score to PROMIS PF.

In conclusion, optimizing the use of PROMs in THA and TKA contributes to improved shared decision-making, outcome assessment, quality of care, and VBHC delivery. Addressing current limitations in measurement validity, reliability, and implementation is critical to realize the full potential of PROMs in orthopedic practice, research, and policy-making. Improving their implementation and addressing current limitations are crucial steps towards achieving patient-centered healthcare in THA and TKA, while having the potential to contribute to environmental sustainability and reducing workforce pressures. Insufficient implementation or inadequate PROMs can lead to inefficient and less patient-centered healthcare.

Aims of the thesis

This thesis aims to optimize the measurement of patient-reported outcomes evaluating physical function and pain in THA and TKA patients.

Outline of the thesis

Part I. Measurement properties of legacy PROMs evaluating physical function in THA and TKA

Part 1 consists of two chapters evaluating the psychometric properties of legacy PROMs measuring physical function in THA and TKA. Chapter 2 systematically reviewed the measurement properties of the HOOS-PS and the KOOS-PS. These instruments were, at the time of the study, recommended by the NOV for evaluating THA and TKA. Chapter 3 evaluated the content validity of the HOOS-PS through interviews with patients and clinicians.

Part II: Towards an adequate alternative patient-reported outcome measure in THA and TKA

Part 1 showed the need for alternative measurement instruments; therefore, Part 2 assesses an innovative alternative. This part consists of two chapters comparing the psychometric properties of the Patient-Reported Outcomes Measurement Information System (PROMIS®) with legacy PROMs that evaluate physical function and pain. Chapter 4 assessed and compared the measurement properties of both PROMs in THA patients. Chapter 5 assessed the measurement properties of both PROMs in patients undergoing TKA.

Part III: Standardizing legacy PROM score conversions towards PROMIS scores

In this part, the existing crosswalks between legacy PROMs in THA and TKA patients towards PROMIS instruments were evaluated. Chapter 6 is a multicenter external validation study evaluating the applicability of available crosswalks at the group and individual level. This contributes to standardizing PROM score conversions and facilitates data continuity in the event of a transition to PROMIS measures.

Discussion

A framework for enhancing PROM utilization in orthopedic care is presented in the Discussion.

References

1. Price AJ, Alvand A, Troelsen A, et al. Knee replacement. *Lancet*. 2018;392(10158):1672-1682. doi:10.1016/S0140-6736(18)32344-4
2. Ferguson RJ, Palmer AJ, Taylor A, Porter ML, Malchau H, Glyn-Jones S. Hip and knee replacement 1 - Hip replacement. *Lancet*. 2018;392.
3. *Online LROI Annual Report 2025*.; 2025. https://www.lroi.nl/media/f3znilzr/pdf-lroi-report-2025-3.pdf.
4. van Veghel MHW, van Steenbergen LN, Gademan MGJ, van den Hout WB, Schreurs BW, Hannink G. How many people in the Netherlands live with a hip, knee, or shoulder replacement? *Bone Jt Open*. 2025;6(1):74-81. doi:10.1302/2633-1462.61.BJO-2024-0162.R1
5. Gossec L, Hawker G, Davis AM, et al. OMERACT/OARSI initiative to define states of severity and indication for joint replacement in hip and knee osteoarthritis. *J Rheumatol*. 2007;34(6):1432-1435. http://www.ncbi.nlm.nih.gov/pubmed/17552070.
6. Smith TO, Hawker GA, Hunter DJ, et al. The OMERACT-OARSI Core Domain Set for Measurement in Clinical Trials of Hip and/or Knee Osteoarthritis. *J Rheumatol*. 2019;46(8):981-989. doi:10.3899/jrheum.181194
7. Dahlberg LE. ICHOM Standard Set for monitoring knee and hip osteoarthritis. *Osteoarthr Cartil*. 2016;24:S436-S437. doi:10.1016/j.joca.2016.01.791
8. Poolman RW, Caron JJ, van de Groes SAW, Pieter van de Ree LC, Pijls BG. World Report: A Snapshot of Orthopaedic Surgery in The Netherlands. *J Bone Jt Surg*. 2025;107(22):2517-2520. doi:10.2106/JBJS.25.00919
9. Proms-advies MNO V. NOV PROMs-advies orthopedie 2020. https://www.orthopeden.org/downloads/775/nov-proms-advies.pdf. Published 2020.
10. Herdman M, Gudex C, Lloyd A, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res*. 2011;20(10):1727-1736. doi:10.1007/s11136-011-9903-x
11. Davis AM, Perruccio A V., Canizares M, et al. The development of a short measure of physical function for hip OA HOOS-Physical Function Shortform (HOOS-PS): an OARSI/OMERACT initiative. *Osteoarthr Cartil*. 2008;16(5):551-559. doi:10.1016/j.joca.2007.12.016
12. Perruccio A V., Stefan Lohmander L, Canizares M, et al. The development of a short measure of physical function for knee OA KOOS-Physical Function Shortform (KOOS-PS) - an OARSI/OMERACT initiative. *Osteoarthr Cartil*. 2008;16(5):542-550. doi:10.1016/j.joca.2007.12.014
13. Wong LH, Meeker JE. The promise of computer adaptive testing in collection of orthopaedic outcomes: an evaluation of PROMIS utilization. *J Patient-Reported Outcomes*. 2022;6(1). doi:10.1186/s41687-021-00407-w
14. Bernstein DN, Fear K, Mesfin A, et al. Patient-reported outcomes use during orthopaedic surgery clinic visits improves the patient experience. *Musculoskeletal Care*. 2019;17(1). doi:10.1002/msc.1379
15. Gibbons C, Porter I, Gonçalves-Bradley DC, et al. Routine provision of feedback from patient-reported outcome measurements to healthcare providers and patients in clinical practice. *Cochrane Database Syst Rev*. 2021;2021(10). doi:10.1002/14651858.CD011589.pub2
16. Makhni EC, Hennekes ME, Baumhauer JF, Muh SJ, Spindler K. AOA Critical Issues: Patient-Reported Outcome Measures. *J Bone Jt Surg*. 2023;105(8):641-648. doi:10.2106/JBJS.22.00587
17. Makhni EC. Meaningful Clinical Applications of Patient-Reported Outcome Measures in Orthopaedics. *J Bone Jt Surg*. 2021;103(1):84-91. doi:10.2106/JBJS.20.00624
18. ViaSana Kliniek. Patients like me. https://www.viasana.nl/behandelingen/persoonlijke-prognose-check/.
19. Geilen JEJ., Hoelen TA, Schotanus MGM, Spekenbrink-Spooren A, Boonen B, Most J. Defining Clinically Meaningful Thresholds for 12-Month Patient-Reported Outcomes in Total Hip Arthroplasty; Toward Improving Threshold Accuracy. *Arthroplast Today*. 2025;32.
20. Pronk Y, Pilot P, van der Weegen W, Brinkman J-M, Schreurs BW. A Patient-Reported Outcome Tool to Triage Total Hip Arthroplasty Patients to Hospital or Video Consultation: Pilot Study With Expert Panels and a Cohort of 1228 Patients. *JMIR Form Res*. 2021;5(12):e31232. doi:10.2196/31232
21. Makhni EC, Hennekes ME. The Use of Patient-Reported Outcome Measures in Clinical Practice and Clinical Decision Making. *J Am Acad Orthop Surg*. 2023;31(20):1059-1066. doi:10.5435/JAAOS-D-23-00040

22. Lavallee DC, Chenok KE, Love RM, et al. Incorporating patient-reported outcomes into health care to engage patients and enhance care. *Health Aff*. 2016;35(4). doi:10.1377/hlthaff.2015.1362
23. Cella D, Hahn E, Jensen S, et al. *Patient-Reported Outcomes In Performance Measurement*. RTI Press; 2015. doi:10.3768/rtipress.2015.bk.0014.1509
24. Porter ME, Olmsted Teisberg E. *Redefining Health Care. Creating Value-Based Competition on Results*.; 2005.
25. Hung IY, Kain ZN, Bozic KJ. Revitalizing Musculoskeletal Healthcare: A Strategic Approach to Value-Based Care. *J Arthroplasty*. 2025;40(2):263-267. doi:10.1016/j.arth.2024.11.026
26. Kuhrij LS, Marang-van de Mheen PJ, van Lier L, Alimahomed R, Nelissen RGHH, van Bodegom-Vos L. Reduction in use of MRI and arthroscopy among patients with degenerative knee disease in independent treatment centers versus general hospitals: a time series analysis. *Int J Qual Heal Care*. 2024;36(1). doi:10.1093/intqhc/mzae004
27. Sociaal Economische Raad. *Zorg Voor de Toekomst Over de Toekomstbestendigheid van de Zorg*.; 2020.
28. RIVM. *Volksgezondheid Toekomst Verkenning (VTV) 2024*.; 2024.
29. Jones CH, Dolsten M. Healthcare on the brink: navigating the challenges of an aging society in the United States. *npj Aging*. 2024;10(22).
30. Braaksma C, Wolterbeek N, Veen MR, Prinsen CAC, Ostelo RWJG. Systematic review and meta-analysis of measurement properties of the Hip disability and Osteoarthritis Outcome Score - Physical Function Shortform (HOOS-PS) and the Knee Injury and Osteoarthritis Outcome Score - Physical Function Shortform (KOOS-PS). *Osteoarthr Cartil*. 2020;28(12):1525-1538. doi:10.1016/j.joca.2020.08.004
31. Braaksma C, Wolterbeek N, Veen RMR, Prinsen CAC, Ostelo RWJG. The Hip Disability and Osteoarthritis Outcome Score-Physical Function Shortform Does Not Adequately Represent Physical Functioning in Patients Undergoing Total Hip Arthroplasty. *Value Heal*. 2022;25(11). doi:10.1016/j.jval.2022.06.001
32. Dawson J, Fitzpatrick R, Carr A, Murray D. Questionnaire on the perceptions of patients about total hip replacement. *J Bone Jt Surg - Ser B*. 1996;78(2):185-190. doi:10.1302/0301-620x.78b2.0780185
33. Dawson J, Fitzpatrick R, Murray D, Carr A. Questionnaire on the perceptions of patients about total knee replacement. *J Bone Jt Surg*. 1998;80(1):63-69. doi:10.1302/0301-620X.80B1.7859
34. Gagnier JJ, Mullins M, Huang H, et al. A Systematic Review of Measurement Properties of Patient-Reported Outcome Measures Used in Patients Undergoing Total Knee Arthroplasty. *J Arthroplasty*. 2017. doi:10.1016/j.arth.2016.12.052
35. Gagnier JJ, Huang H, Mullins M, et al. Measurement properties of patient-reported outcome measures used in patients undergoing total hip arthroplasty: A systematic review. *JBJS Rev*. 2018. doi:10.2106/JBJS.RVW.17.00038
36. Holmenlund C, Overgaard S, Bilberg R, Varnum C. Evaluation of the Oxford Hip Score: Does it still have content validity? Interviews of total hip arthroplasty patients. *Health Qual Life Outcomes*. 2021;19(1). doi:10.1186/s12955-021-01869-8
37. Horn ME, Reinke EK, Mather RC, O'Donnell JD, George SZ. Electronic health record–integrated approach for collection of patient-reported outcome measures: a retrospective evaluation. *BMC Health Serv Res*. 2021;21(1):626. doi:10.1186/s12913-021-06626-7
38. Harris IA, Peng Y, Cashman K, et al. Association between patient factors and hospital completeness of a patient-reported outcome measures program in joint arthroplasty, a cohort study. *J Patient-Reported Outcomes*. 2022;6(1):32. doi:10.1186/s41687-022-00441-2
39. Prodinger B, Taylor P. Improving quality of care through patient-reported outcome measures (PROMs): expert interviews using the NHS PROMs Programme and the Swedish quality registers for knee and hip arthroplasty as examples. *BMC Health Serv Res*. 2018;18(1):87. doi:10.1186/s12913-018-2898-z
40. Brodke DJ, Hung M, Bozic KJ. Item Response Theory and Computerized Adaptive Testing for Orthopaedic Outcomes Measures. *J Am Acad Orthop Surg*. 2016;24(11):750-754. doi:10.5435/JAAOS-D-15-00420
41. Cappelleri JC, Jason Lundy J, Hays RD. Overview of Classical Test Theory and Item Response Theory for the Quantitative Assessment of Items in Developing Patient-Reported Outcomes Measures. *Clin Ther*. 2014;36(5):648-662. doi:10.1016/j.clinthera.2014.04.006
42. Thomas ML. The Value of Item Response Theory in Clinical Assessment: A Review. *Assessment*. 2011;18(3):291-307. doi:10.1177/1073191110374797

43. Cella D, Riley W, Stone A, et al. The patient-reported outcomes measurement information system (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. *J Clin Epidemiol*. 2010;63(11):1179-1194. doi:10.1016/j.jclinepi.2010.04.011
44. Ziedas AC, Abed V, Swantek AJ, et al. Patient-Reported Outcomes Measurement Information System (PROMIS) Physical Function Instruments Compare Favorably With Legacy Patient-Reported Outcome Measures in Upper- and Lower-Extremity Orthopaedic Patients: A Systematic Review of the Literature. *Arthrosc J Arthrosc Relat Surg*. 2022;38(2):609-631. doi:10.1016/j.arthro.2021.05.031
45. Abma IL, Butje BJD, ten Klooster PM, van der Wees PJ. Measurement properties of the Dutch–Flemish patient-reported outcomes measurement information system (PROMIS) physical function item bank and instruments: a systematic review. *Health Qual Life Outcomes*. 2021;19(1):62. doi:10.1186/s12955-020-01647-y
46. Oude Voshaar MAH, ten Klooster PM, Glas CAW, et al. Validity and measurement precision of the PROMIS physical function item bank and a content validity–driven 20-item short form in rheumatoid arthritis compared with traditional measures. *Rheumatology*. July 2015:kev265. doi:10.1093/rheumatology/kev265
47. Czerwonka N, Gupta P, Desai SS, Hickernell TR, Neuwirth AL, Trofa DP. Patient-reported outcomes measurement information system instruments in knee arthroplasty patients: a systematic review of the literature. *Knee Surg Relat Res*. 2023;35(1):27. doi:10.1186/s43019-023-00201-6
48. Kagan R, Anderson MB, Christensen JC, Peters CL, Gililland JM, Pelt CE. The Recovery Curve for the Patient-Reported Outcomes Measurement Information System Patient-Reported Physical Function and Pain Interference Computerized Adaptive Tests After Primary Total Knee Arthroplasty. *J Arthroplasty*. 2018;33(8):2471-2474. doi:10.1016/j.arth.2018.03.020
49. Braaksma C, Wolterbeek N, Veen MR, et al. Assessing the measurement properties of PROMIS Computer Adaptive Tests, short forms and legacy patient reported outcome measures in patients undergoing total hip arthroplasty. *J Patient-Reported Outcomes*. 2024;8(1):121. doi:10.1186/s41687-024-00799-5
50. Zonjee VJ, Abma IL, de Mooij MJ, et al. The patient-reported outcomes measurement information systems (PROMIS®) physical function and its derivative measures in adults: a systematic review of content validity. *Qual Life Res*. 2022;31(12):3317-3330. doi:10.1007/s11136-022-03151-w
51. Terwee C, Ahmed S, Alhasani R, et al. Comparable real-world patient-reported outcomes data across health conditions, settings, and countries: the PROMIS international collaboration. *NEJM Catal*. 2024;5(9).
52. Cella MS, Baumhauer JF, Rothrock NE, Swantek K, Franklin PD. Use of Patient-Reported Outcomes Measurement Information System Measures in Orthopaedic Specialties: Results of a Scoping Review for 2018 to 2022. *J Am Acad Orthop Surg*. 2025;33(11):561-568. doi:10.5435/JAAOS-D-24-00432
53. Stephan A, Stadelmann VA, Preiss S, Impellizzeri FM. Measurement properties of PROMIS short forms for pain and function in patients receiving knee arthroplasty. *J Patient-Reported Outcomes*. 2023;7(1):18. doi:10.1186/s41687-023-00559-x
54. Stephan A, Mainzer J, Kümmel D, Impellizzeri FM. Measurement properties of PROMIS short forms for pain and function in orthopedic foot and ankle surgery patients. *Qual Life Res*. 2019;28(10):2821-2829. doi:10.1007/s11136-019-02221-w
55. Schalet BD, Lim S, Cella D, Choi SW. Linking Scores with Patient-Reported Health Outcome Instruments:A VALIDATION STUDY AND COMPARISON OF THREE LINKING METHODS. *Psychometrika*. 2021;86(3). doi:10.1007/s11336-021-09776-z
56. Tang X, Schalet BD, Heng M, et al. Linking the KOOS-PS to PROMIS Physical Function in Knee Patients Evaluated for Surgery. *J Am Acad Orthop Surg*. 2022;30(6). doi:10.5435/JAAOS-D-21-00461
57. Heng M, Stern BZ, Tang X, et al. Linking Hip Disability and Osteoarthritis Outcome Score-Physical Function Short Form and PROMIS Physical Function. *J Am Acad Orthop Surg*. 2022;30(15). doi:10.5435/JAAOS-D-21-00736
58. Crins MHP, van der Wees PJ, Klausch T, van Dulmen SA, Roorda LD, Terwee CB. Psychometric properties of the PROMIS Physical Function item bank in patients receiving physical therapy. *PLoS One*. 2018;13(2). doi:10.1371/journal.pone.0192187
59. Crins MHP, Terwee CB, Klausch T, et al. The Dutch–Flemish PROMIS Physical Function item bank exhibited strong psychometric properties in patients with chronic pain. *J Clin Epidemiol*. 2017;87. doi:10.1016/j.jclinepi.2017.03.011

PART I

Measurement properties of legacy PROMs evaluating physical function in THA and TKA

CHAPTER 2

Systematic review and meta-analysis of measurement properties of the Hip disability and Osteoarthritis Outcome Score – Physical Function Shortform (HOOS-PS) and the Knee Injury and Osteoarthritis Outcome Score - Physical Function Shortform (KOOS-PS)

Osteoarthritis and Cartilage 2020 Dec; 28(12):1525-1538

Abstract

Objective

The aim of this systematic review and meta-analysis was to evaluate all evidence on measurement properties of the Hip disability and Osteoarthritis Outcome Score - Physical function Shortform (HOOS-PS) and the Knee Injury and Osteoarthritis Outcome Score - Physical function Shortform (KOOS-PS).

Design

This study was conducted according to the COSMIN guideline for systematic reviews of PROMs. MEDLINE, EMBASE, The Cochrane Library, CINAHL and PsychINFO through February 2019 were searched. Eligible studies evaluated patients with hip or knee complaints and described a measurement property, interpretability, feasibility, or the development of either the HOOS-PS or KOOS-PS.

Results

Twenty-three studies were included. For both questionnaires, the content validity was found inconsistent and the quality evidence was moderate for a sufficient reliability and high for an insufficient construct validity. The HOOS-PS had a high quality evidence of sufficient structural validity and internal consistency (pooled Cronbach's alpha 0.80; n=3761) and low quality evidence of sufficient measurement error and indeterminate responsiveness. Concerning the KOOS-PS, the quality evidence was high for an insufficient responsiveness, moderate for an inconsistent structural validity and internal consistency and low for an inconsistent measurement error.

Conclusions

The inconsistent evidence for content validity implies that scores on the HOOS-PS and KOOS-PS may inadequately reflect physical functioning. Furthermore, there is evidence for insufficient construct validity and responsiveness in patients with knee osteoarthritis receiving conservative treatment. Using the HOOS-PS or KOOS-PS as outcome measurement instruments for comparing outcomes, measuring improvements or benchmarking in patients with hip or knee complaints or undergoing arthroplasty should only be done with great caution.

Introduction

In total joint arthroplasty, patient reported outcome measures (PROMs) are widely used to evaluate the effect of treatment on individual patients and for comparative effectiveness research. In addition, the health care industry has become interested in using these instruments as an indicator of quality of care[1].

Widely used PROMs measuring physical functioning in patients with hip or knee complaints are the Hip disability and Osteoarthritis Outcome Score - Physical function Shortform (HOOS-PS)[2] and the Knee Injury and Osteoarthritis Outcome Score - Physical function Shortform (KOOS-PS)[3], respectively. The items on the HOOS-PS and KOOS-PS were selected using Rasch analysis of the Western Ontario McMaster Universities Osteoarthritis Index (WOMAC)[4] and the full-length HOOS[5] and KOOS[6]. The HOOS-PS and KOOS-PS aim to measure physical functioning with fewer items and similar validity compared to the full-length measurements instruments, in order to minimize the burden of the responder and decrease the administrative load. The HOOS-PS and KOOS-PS are selected as outcome measurement instruments by global standard sets of outcome measures, arthroplasty registries and clinical research studies[7–9].

Although the full-length HOOS and KOOS are extensively evaluated, the measurement properties of the short forms of these questionnaires have not been summarized[10–13]. The available systematic reviews did not pool the data quantitatively, included only one article or did not focus on the short form measurement instruments[10–13]. Furthermore, the PROM development and content validity were not qualitatively evaluated. It is important to assess if the HOOS-PS and KOOS-PS are a valid reflection of physical functioning since the outcomes of these measurement instruments are used to evaluate individual patients and to benchmark health care providers.

The goal of this systematic review and meta-analysis is to evaluate all evidence on the measurement properties (content validity, structural validity, internal consistency, reliability, measurement error, cross-cultural validity/measurement invariance, construct validity, criterion validity and responsiveness) and the interpretability of the HOOS-PS and KOOS-PS in patients with hip or knee complaints or undergoing total hip or knee arthroplasty.

Materials and methods

Protocol and registration

This review was reported according to the Preferred Reporting Items for Systematic review and Meta-Analysis Protocols (PRISMA-p)[14]. A study protocol was registered in PROSPERO [CRD42017069539]. The systematic review was conducted according to the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) guideline for systematic reviews of PROMs[15]. COSMIN aims to improve the selection of outcome measurement instruments by developing methodology and practical tools for selecting the most suitable outcome measurement instrument.

Eligibility criteria

Eligible studies were full text articles evaluating at least one measurement property or the interpretability of the HOOS-PS and KOOS-PS, or reporting on the development of either the HOOS-PS or KOOS-PS. Furthermore, the development studies of the WOMAC, full-length HOOS or KOOS were eligible, since the items of the HOOS-PS and KOOS-PS were extracted from these measurement instruments in unchanged form. All studies had to evaluate patients of any age with hip or knee complaints or patients who underwent arthroplasty. Included measurement properties were the content validity, structural validity, internal consistency, reliability, measurement error, cross-cultural validity/measurement invariance, construct validity, criterion validity and responsiveness. Table 2.1 provides an overview of the definitions of the measurement properties and the interpretability. The HOOS-PS and KOOS-PS had to be patient reported or research administrator assisted. Reviews, study protocols or studies using the outcome measurement instruments for assessment of patients with other limb conditions than hip or knee complaints were excluded. The search was not restricted on language, publication status or study design.

Searches

A literature search was performed in the following electronic bibliographic databases (February 11, 2019): MEDLINE through PubMed, EMBASE through OVID, The Cochrane Library (Cochrane Database of Systematic Reviews, Cochrane Central Register of Controlled Trials, Cochrane Methodology Register), CINAHL and PsychINFO. The search strategy was reviewed by a clinical librarian and can be found in the Supplemental material. References were searched manually to identify other potential studies. Furthermore, the website http://www.koos.nu was checked for other publications or PhD theses.

Data extraction (selection and coding)

After duplicate removal, two reviewers (CB and NW) identified potentially eligible studies after assessing title and abstract of the retrieved studies independently. If one or both of the reviewers identified a study as potentially eligible, the full text was retrieved and independently assessed by the same two reviewers (CB and NW). Studies were included if they met the eligibility criteria.

The data from the included studies were extracted using a data extract template of the COSMIN manual for systematic reviews of PROMS[16]. The extraction was done by one reviewer (CB) and the second reviewer checked the extracted data (NW) on patient characteristics (number of participants, mean age, sex distribution, disease characteristics, response rate) and type of measurement (HOOS-PS, KOOS-PS, time interval used for follow-up, setting in which the study was conducted, country, language, mode of administration) and all information available on measurement properties. In case differences, consensus was reached by discussion.

Table 2.1. Taxonomy of the measurement properties and the interpretability, obtained from Mokkink et al (2010).[25]

Measurement property	Definition
Content validity	The degree to which the content is an adequate reflection of the construct to be measured
Structural validity	The degree to which the scores are an adequate reflection of the dimensionality of the construct to be measured
Internal consistency	The degree of the interrelatedness among the items
Reliability	The proportion of the total variance in the measurements which is because of true differences among patients
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
Cross-cultural validity / measurement invariance	The degree to which the performance of the items on a translated or culturally adapted PROM are an adequate reflection of the performance of the items of the original version
Construct validity	The degree to which the scores are consistent with hypotheses based on the assumption that the PROM validly measures the construct to be measured
Criterion validity	The degree to which the scores are an adequate reflection of a ''gold standard''
Responsiveness	The ability to detect change over time in the construct to be measured
Interpretability	The degree to which one can assign qualitative meaning (that is, clinical or commonly understood connotations) to a PROM's quantitative scores or change in scores.

Strategy for data synthesis

The methodological quality of the identified studies was assessed per measurement property (taxonomy of measurement properties, Table 2.1) according to the recently updated COSMIN Risk of Bias checklist[17]. Per study, the methodological quality of the measurement property was scored by two independent authors (CB and NW) on a four-point rating scale (i.e. 'very good', 'adequate', 'doubtful' or 'inadequate' quality)[18]. Subsequently, each measurement property was evaluated against the criteria for good measurement properties per study as 'sufficient', 'insufficient' or 'indeterminate'[15]. The quality criteria for good measurement properties are available in the Supplemental materials. A third reviewer was consulted if no consensus was reached (CP).

Summarize quality of evidence and pooling evidence

The overall quality of the PROM was determined using the modified GRADE approach[15]. taking into account the methodological quality of the studies and the quality of the measurement properties. The modified GRADE approach was used to downgrade the quality of evidence when there are concerns regarding the risk of bias (evaluated by the COSMIN Risk of Bias checklist), inconsistency in results, imprecision and indirect results[15]. The modified GRADE approach is described in detail in the COSMIN manual for systematic reviews[16]. Quality was graded as 'high', 'moderate', 'low' or 'very low'. The evidence on the measurement properties was pooled quantitatively when the studies were comparable in terms of study population and methodological quality. Otherwise, they were qualitatively summarized. To be able to pool the results of the construct validity and the responsiveness, the authors defined hypotheses about the expected correlations between the HOOS-PS or KOOS-PS and comparator instruments (Table 2.2). All correlations of the (changes in) HOOS-PS and KOOS-PS scores with the comparator instruments found in the included studies were tested against the predefined hypotheses. Afterwards, the percentage of accepted hypotheses and the studies were pooled by calculating the weighted average of the correlations. Discrepancies regarding the pooling of the results and grading of the evidence were resolved by discussion. A third reviewer was consulted when needed (CP).

Table 2.2. Predefined hypotheses: the expected correlations between the HOOS-PS or KOOS-PS and comparator instruments.

Number	Hypothesis
1	Correlations with (changes in) instruments measuring physical function like the physical function subscale of the WOMAC, the KOOS/HOOS and the Oxford Hip Score (OHS)/Oxford Knee Score (OKS) should be >0.50
2	Correlations with (changes in) instruments measuring pain (like the pain subscale of either the WOMAC, OKS/OHS or KOOS/HOOS) or stiffness (like the WOMAC stiffness subscale) should be 0.30-0.50
3	Correlations with (changes in) instruments measuring unrelated constructs like mental health or social functioning should be <0.30
4	Correlations with (changes in) instruments measuring similar constructs should differ by a minimum of 0.10 from correlations with (changes in) instruments measuring related but dissimilar constructs
5	Correlations with (changes in) instruments measuring related constructs should differ by a minimum of 0.10 from correlations with (changes in) instruments measuring unrelated constructs

Statistical analysis

Meta-analysis was done following the method of Feldt and Charter (2006) to compute the pooled internal consistency[19]. Cronbach's alphas were transformed to Fisher's Ƶ values that were averaged (weighted average for sample size per study) and converted back to a pooled Cronbach's alpha. Stepwise approach:

1. Calculate a Ƶ value per Cronbach's alpha[19]
   $Ƶ = 1.1513 \{\log_{10} [(1 + r) / (1 - r)]\}$
2. Calculate the average weighted Ƶ[19]
   $\bar{ƶ} = \sum (n_j - 3)ƶ_j / \sum (n_j - 3)$
3. Convert the Ƶ value back to a pooled Cronbach's alpha[19]
   $r = (10^{ƶ/1.1513} - 1) / (10^{ƶ/1.513} + 1)$

We combined the framework of DerSimonian (1986)[20] and Feldt and Charter (2006)[19] to compute the pooled test-retest reliability. Fisher's transformation to Ƶ values were computed by the method of DerSimonian(20). Computing weighted average was done for the ICC and the confidence interval (95%) the same as for the Cronbach's alpha, with the method of Feldt[19]. Stepwise approach:

1. Calculate the ƶ value per ICC[20]
   ƶ = 0.5 x ln ((1 + ICC) / (1 - ICC))
2. Calculate the average weighted ƶ[19]
   $\bar{ƶ} = \Sigma (n_j - 3)ƶ_j / \Sigma (n_j - 3)$
3. Convert the ƶ value back to a pooled ICC[19]
   $r = (10^{ƶ/1.1513} - 1) / (10^{ƶ/1.513} + 1)$

Results

The results of the literature search and selection of the studies are displayed in the PRISMA flow diagram (Figure 2.1). The characteristics of the included PROMs are presented in Table 2.3. The characteristics of the included studies and their populations are presented in Table 2.4. The summary of findings for each measurement property is presented in Table 2.5.

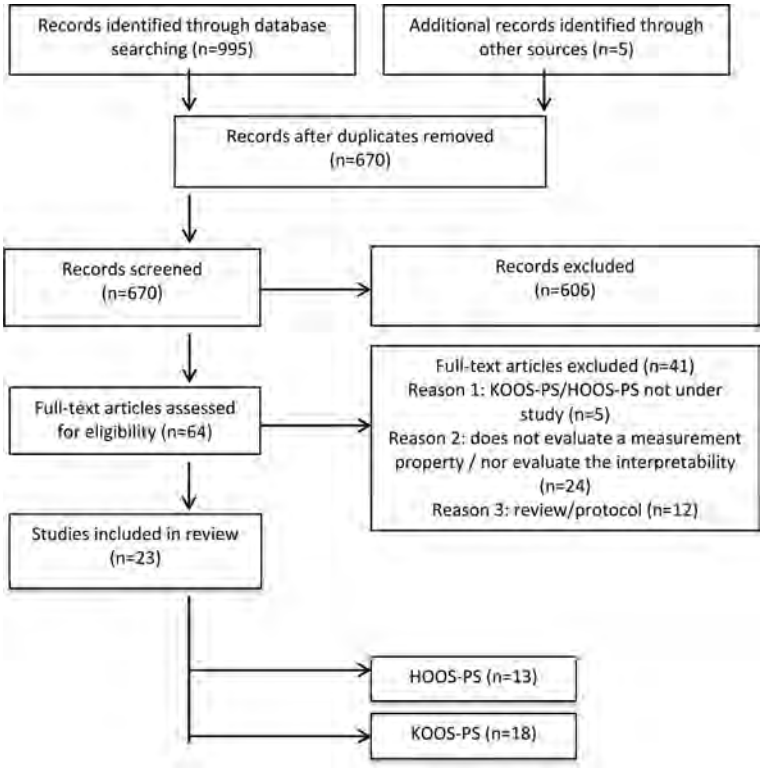

Figure 2.1. PRISMA flow diagram of the literature search and selection of the studies.

Content validity

The way PROMs are developed effects the content validity. The HOOS-PS and KOOS-PS were developed via Rasch analysis of the full-length HOOS, KOOS and WOMAC and tested in populations of all ages, from several countries with a wide spectrum of severity of osteoarthritis. The construct to be measured and the target population were clearly described[2,3]. However, no theoretical framework was used to define the construct in a

broader setting. The items of the outcome measurement instruments were created in the development studies of the full-length versions and selected in unchanged form. Therefore, the methodology and possible limitations of the PROM development studies of the full-length HOOS, KOOS and WOMAC effect the methodological quality of the shorter versions. The items in the development studies were created based on literature review, consulting expert panels and pilot studies[4–6]. No cognitive interviews were conducted to evaluate the comprehensiveness or comprehensibility.

Table 2.3. Characteristics of the included PROMs.

Measurement property	HOOS-PS(2)	KOOS-PS(3)
Construct	Physical function	Physical function
Target population	People with hip problems	People with knee problems
Mode of administration	Self-administered	Self-administered
Recall period	1 week	1 week
Scale (number of items)	1 (5)	1 (7)
Response options	None / Mild / Moderate / Severe / Extreme	None / Mild / Moderate / Severe / Extreme
Range of scores/scoring	0-100 (with 0 representing extreme difficulty)	0-100 (with 0 representing extreme difficulty)
Original language	English	English
Available translations	Danish Dutch[1] English, French[1] German Italian Norwegian Polish Portuguese (Brazil) Swedish Turkish[1]	Arabic (Saudi Arabia) Chinese Danish Dutch[1] English French[1] German Hindi (India) Italian Korean Norwegian Polish Portuguese[1] Portuguese (Brazil) Singapore (English) Spanish Swedish Turkish[1]

1: Validated translations

The content validity was inconsistent of the HOOS-PS[2,4,5,21,22] and KOOS-PS[3,4,21,23,24]. Content validity refers to the relevance (the degree to which the content is considered applicable for measuring physical functioning), comprehensiveness (the degree to which all key aspects of the constructs are covered) and the comprehensibility (the degree to which the items, response options and instructions are understood by patients as intended)[25]. Consecutively the relevance, comprehensiveness and comprehensibility are discussed.

Table 2.4. Characteristics of the included study populations.

Reference	Country; evaluated language	Setting	single-centre or multi-centre study	PROM	Number of patients	Age in years Mean(SD), range	Diagnosis	Gender	FU
Bond 2012[38]	USA; English	Interview-administered	multicentre	HOOS-PS	48	60.3(9.4)	Hip OA, conservative treatment	68.8% female	13 weeks
				KOOS-PS	156	61.2(9.2)	Knee OA, conservative treatment	68.8% female	13 weeks
Davis 2009[28]	Canada; English	Patient-administered, setting unclear	multicentre	HOOS-PS	201	62.3(12.1)	Hip OA, pre and post THR	53% female	6 months
				KOOS-PS	248	64.5(10.3)	Knee OA, pre and post TKR	63% female,	6 months
Davis 2008[2]	Countries: Canada, Sweden, Austria, Finland, France, Germany, Hungary, Iceland, Italy, Poland, Spain, Sweden, Switzerland, United Kingdom; multiple languages.	Patient-administered, setting unclear	multicentre	HOOS-PS	2991	range 19-96	pre-THR surgery cohorts, community cohort	male: female 1:1.23	NA
Franchignoni 2013[26]	Italy; Italian	Patient-administered, setting unclear	single centre	KOOS-PS	200	69.4(9.5), range 50-84	Knee OA	73.5% female	NA
Goncalves 2010[31]	Portugal; Portuguese	Patient-administered, setting unclear	multicentre	KOOS-PS	85	65.7(6.9)	Knee OA	74.1% female	48 hour, 4 weeks or 6 weeks
Gul 2013[32]	Turkey; Turkish	Unclear setting and administration	single centre	KOOS-PS	80	58.9(8.7), range 42-76	Knee OA	88.7% female	NA

Table 2.4. (continued).

Reference	Country; evaluated language	Setting	single-centre or multi-centre study	PROM	Number of patients	Age in years Mean(SD), range	Diagnosis	Gender	FU
Harris 2013[27]	England; English	Patient-administered paper by mail	single centre	KOOS-PS	134	59(11)	Knee OA	50% female	3 months
Mahler 2016[41]	Netherlands; Dutch	Patient-administered paper by mail	single centre	KOOS-PS	161	59(9)	Knee OA	61% female	3 months
Mehta 2016[29]	Sweden, UK, Australia Canada, Czech republic, France, Netherlands; multiple languages	Patient-administered, setting unclear	multicentre	HOOS-PS KOOS-PS	745 1064	64.9(11.4) 66.8(10.6)	Hip OA Knee OA	57% female 58% female	NA NA
Ornetti 2009[36]	France; French	Baseline unclear, follow-up patient-administered setting unclear	single centre	HOOS-PS KOOS-PS	50 87	65(10) 72(9)	Hip OA Knee OA	74% female 71% female	Up to 1 month Up to 1 month
Paulsen 2014[37]	Denmark; Danish	Baseline unclear, follow-up patient-administered paper by mail	multicentre	HOOS-PS	1335	68, range 23-94	1175 hip OA, 45 other arthritis, 30 childhood hip diseases, 6 sequel from fracture, 7 necrosis of femoral head	54% female	1 year
Perrucio 2008[3]	Sweden, Canada, France, Estonia, Netherlands; multiple languages	Patient-administered, setting unclear	multicentre	KOOS-PS	2145	range 26-95	community, knee OA, medial wedge, pre-osteotomy, post ACL	male: female 1:1.4	NA

Table 2.4. (continued).

Reference	Country; evaluated language	Setting	single-centre or multi-centre study	PROM	Number of patients	Age in years Mean(SD), range	Diagnosis	Gender	FU
Ruyssen 2011[30]	France; French	Patient-administered, setting unclear	single centre	HOOS-PS	172 validity 33 reliability 107 responsiveness	65.1(12.3) 64.7(12.1) 65.6(10.2)	Hip OA	53.5% female 63.6% female 48.6% female	12 weeks
				KOOS-PS	128 validity 30 reliability 60 responsiveness	70.9 (10.5) 69.3 (10.9) 71 (10.3)	Knee OA	72.7% female 66.7% female 68.3% female	12 weeks
Singh 2014[35]	USA; English	Patient-administered, setting unclear	multicentre	HOOS-PS	54	Hip and knee cohort together: 60.8 (11.4)	Hip OA	Hip and knee cohort together: 42.6% female	Up to 20 days
				KOOS-PS	141	Hip and knee cohort together: 60.8 (11.4)	Knee OA	Hip and knee cohort together: 42.6% female	Up to 20 days
Stratford 2014[40]	Canada; English	Patient-administered, setting unclear	single centre	KOOS-PS	377	64.4(10.5)	Knee OA	63% female	NA
Wiering 2017[21]	Netherlands; Dutch	Patient-administered online and patient-administered paper	multicentre	HOOS-PS	1393	72(9.1)	Post THA	Hip and knee cohort together: 65.7% female	NA
				KOOS-PS	1278	72(9.1)	Post TKA	Hip and knee cohort together: 65.7% female	NA

Table 2.4. (continued).

Reference	Country; evaluated language	Setting	single-centre or multi-centre study	PROM	Number of patients	Age in years Mean(SD), range	Diagnosis	Gender	FU
Yilmaz 2014[34]	Turkey; Turkish	Literate patients: patient-administered, setting unclear. Illiterate patients were read aloud by an investigator.	single centre	HOOS-PS	50	59.1(9.2), range 41-77	Hip OA	74% female	1 week
Groot 2008[23]	Netherlands; Dutch	Patient-administered, setting unclear	unclear	KOOS-PS	15	unclear	Knee OA	unclear	NA
Groot 2007[22]	Netherlands; Dutch	Patient-administered, setting unclear	unclear	HOOS-PS	15	unclear	Hip OA	unclear	NA
Roos 1998[24]	Sweden, USA; Swedish and English	Patient-administered, setting unclear	unclear	KOOS	75	56, range 35-76	Knee OA	not described	NA
Klassbo 2003[5]	Sweden; Swedish	Patient-administered, setting unclear	single centre	HOOS	52	64, range 42-48	Hip complaints, patients with and without hip OA	female / male 35/17	NA
Bellamy 1986[4]	Canada; English	90 face to face interview-administered, 10 telephone interview administered, 15 patients unclear	multicentre	WOMAC	100 (11 hip, 57 knee and 32 both hip and knee)	61.07, range 27-93	Hip or knee OA	female / male 63 / 37	NA
Gandek 2019[46]	USA; English	Patient administered, either paper-pencil or on internet, at the outpatient clinic or at home	Multicentre	KOOS-PS	1295	66.5, range 37-100	Knee OA	68.2% female	

Abbreviations: OA: osteoarthritis; TKR: total knee replacement; THR: total hip replacement; NA: not applicable.

Table 2.5. Summary of findings.

Content validity	Summary (methodologic rating)	Overall rating	Quality of evidence
HOOS-PS[2,4,5,21,22]	Inconsistent relevance (very low), insufficient comprehensiveness (very low) and sufficient comprehensibility (moderate). None of the included studies evaluated all domains of content validity.	Inconsistent	No grading, since overall rating was inconsistent it is not possible to judge quality of evidence
KOOS-PS[3,4,21,23,24]	Inconsistent relevance (very low), insufficient comprehensiveness (very low) and sufficient comprehensibility (moderate). None of the included studies evaluated all domains of content validity.	Inconsistent	No grading, since overall rating was inconsistent it is not possible to judge quality of evidence
Internal consistency	**Summary or pooled result**	**Overall rating**	**Quality of evidence**
HOOS-PS[2,28–30]	Pooled Cronbach's alpha = 0.80; total sample size 3761	Sufficient	High as there were several studies with very good methodology
KOOS-PS[26–31,33,34]	Pooled Cronbach's alpha = 0.85; total sample size 3212	Indeterminate	Moderate as the structural validity was inconsistent
Cross-cultural validity	**Summary or pooled result**	**Overall rating**	**Quality of evidence**
HOOS-PS	No info available	No info available	No info available
KOOS-PS	No info available	No info available	No info available
Reliability	**Summary or pooled result**	**Overall rating**	**Quality of evidence**
HOOS-PS[30,34–36]	Pooled ICC = 0.86 (0.67-0.91); total sample size 142	Sufficient	Moderate as there was very serious risk of bias (all studies doubtful methodology)
KOOS-PS[30–32,35,36]	Pooled ICC = 0.81 (0.67-0.87); total sample size 291	Sufficient	Moderate as there was very serious risk of bias (all studies doubtful methodology)

Table 2.5. (continued).

Measurement error	Summary or pooled result	Overall rating	Quality of evidence
HOOS-PS[36,37]	LoA < MIC	Sufficient	Low as there was very serious risk of bias (only one study with doubtful methodology)
KOOS-PS[27,35,37]	Inconsistent results	Indeterminate	Low as there was serious risk of bias (two studies with doubtful methodology) and there were inconsistent results
Hypotheses testing	**Summary or pooled result**	**Overall rating**	**Quality of evidence**
HOOS-PS[28–30,34,36,38]	3 out of 5 results in accordance with hypotheses	Insufficient	High: there were several studies with adequate methodology. As the hypotheses came from inadequate comparator instruments, we ignored these results
KOOS-PS[27–32,36,38,40]	3 out of 5 results in accordance with hypotheses	Insufficient	High: there were several studies with adequate methodology. As the hypotheses came from inadequate comparator instruments, we ignored these results
Responsiveness	**Summary or pooled result**	**Overall rating**	**Quality of evidence**
HOOS-PS[27,28,30,31,36,38,41]	No data available of studies with an adequate methodology	Indeterminate	Very low: as there were only studies with inadequate methodology and there were inconsistent results
KOOS-PS[27,41]	2 out of 5 results in accordance with hypotheses	Insufficient	High: as we included two studies with very good methodology and these had consistent results

Abbreviations: ICC: intraclass correlation; LoA: limit of agreement; MIC: minimally important change.

There was low quality evidence for an inconsistent relevance of the items of the HOOS-PS and KOOS-PS. The full-length HOOS and KOOS development studies evaluated the items on relevance in patients and included the items with the highest responses. However, they did not use a cut-off value for inclusion of the items[5,6]. The relevance of the items of the HOOS-PS and KOOS-PS was determined in patients undergoing hip or knee arthroplasty and was considered insufficient[21]. In this study including more than 1200 patients, the item 'running' of the HOOS-PS was found unimportant by 77.7% of patients[21]. In the same study, the items 'kneeling' and 'squatting' of the KOOS-PS were found unimportant by 32.7% and 39.5% of the patients, respectively[21]. The appropriateness of the response options and recall period, and the relevance for the construct of interest and the context of use were not evaluated.

There was very low quality evidence for an insufficient comprehensiveness of the HOOS-PS and KOOS-PS. As no studies evaluated the comprehensiveness, this rating is based on the reviewers rating solely.

There was moderate quality evidence for a sufficient comprehensibility of the HOOS-PS and KOOS-PS. Evidence regarding the comprehensibility is available from studies translating or developing the items of the full-length HOOS, KOOS and WOMAC[4,5,21–24]. The WOMAC development study evaluated the comprehensibility and relevance of a part of the items of the HOOS-PS and KOOS-PS, however not all items[4]. The translated full-length HOOS and KOOS into Dutch were rated as comprehensible in a sample of 15 patients per study[22,23], however, methodological quality of these studies was doubtful. It is not clear if skilled group interviewers were used or an appropriate interview guide, if the interviews were recorded and transcribed, how the data was evaluated and analysed and if (besides the items) the instructions and response options were evaluated as well.

Structural validity

There was high quality evidence for a sufficient structural validity of the HOOS-PS. The PROM development study (methodological quality rated as 'very good') assessed the structural validity in a sample of 2643 persons.[2] Confirmatory factor analysis (CFA) showed a unidimensional construct and showed there was no clustering (location item mean 0 (SD 1.64), X^2 42.29 with a probability of 0.0672, PSI 0.80).

There was moderate quality evidence for an inconsistent structural validity of the KOOS-PS. The KOOS-PS was developed using a Rasch analysis (methodology rated as 'very good')[3] and showed with CFA that the one factor (unidimensional) structure has an

adequate fit (location item mean 0 (SD 1.229), X^2 73.34 with a probability of 0.1751, PSI 0.904). Two studies repeated the analysis of the items. First, Franchignoni et al (methodology rated as 'adequate') could not replicate the selection of items of the KOOS-PS in patients with knee osteoarthritis[26]. The items "Twisting/pivoting on your injured knee" showed a borderline infit value and "Rising from bed" showed overfit and thus did not fit the Rasch model. Second, Harris et al. 2013 (methodology rated as 'very good') showed with CFA that there was no acceptable evidence to support the structural validity of the KOOS-PS in 113 knee osteoarthritis patients[27].

Internal consistency

There was high quality evidence for a sufficient internal consistency of the HOOS-PS. Pooled Cronbach's alpha in four studies with good methodological quality was 0.80 for the HOOS-PS in 3761 patients[2,28–30].

There was moderate quality evidence for an indeterminate internal consistency of the KOOS-PS. Since we showed that the KOOS-PS is not unidimensional, the internal consistency (pooled outcome of the Cronbach's alpha (0.85 in 3212 patients[26–33], is difficult to interpret and could not be used and the overall rating was scored as indeterminate. One study was excluded from pooling, because of doubtful methodological quality[34]. One study was rated as sufficient after a discussion within the research team, despite of a Cronbach's alpha of 0.69[30].

Reliability

There was moderate quality evidence for a sufficient reliability of the HOOS-PS and KOOS-PS. Pooled ICC of the HOOS-PS was 0.86 (95% CI 0.67-0.91) in 142 patients[30,34–36]. Pooled ICC of the KOOS-PS was 0.81 (95% CI 0.67-0.87) in 291 patients[30–32,35,36]. Major reasons for the moderate quality evidence were the inclusion of less than fifty subjects per study and not being clear if test conditions and the situation of the patients were similar at baseline and retest. One study was rated as sufficient after consensus meeting, despite an ICC of 0.66[35].

Measurement error

There was low quality evidence of a sufficient measurement error of the HOOS-PS. Limits of agreement (LoA)[36] were smaller than the minimally important change (MIC) obtained from another included study[37] so the measurement error was rated sufficient.

There was low quality evidence for an inconsistent measurement error of the KOOS-PS. The measurement error could not be pooled because there were inconsistent results between studies, probably explained by methodological flaws. The first study showed that the standard error of measurement of 6.7 and an anchor based MIC of 12 results in a smallest detectable change of 18.6 points. This is larger than the MIC, so the measurement error was insufficient[27]. The second study showed that the LoA was smaller than the MIC so the rating was sufficient (no absolute numbers available for the LoA, MIC 28 obtained from another study)[35,36].

Cross-cultural validity and measurement invariance

Cross-cultural validity and measurement invariance could not be evaluated, because no studies evaluated this measurement property of either the HOOS-PS or the KOOS-PS.

Hypotheses testing for construct validity

There was high quality evidence for an insufficient construct validity of the HOOS-PS and KOOS-PS. 60% of the results were in accordance with the hypotheses of both the HOOS-PS and KOOS-PS (Table 2.2); this is below the threshold of 75% for a sufficient rating (Table 2.6).

Six studies determined the construct validity of the HOOS-PS by correlations with comparator measurement instruments, containing a total of 20 correlations[28–30,34,36,38]. 60% of the results were in accordance with the hypotheses (3 out of 5).

The construct validity of the KOOS-PS was evaluated in nine studies, with a total of 35 correlations of the KOOS-PS with comparator measurement instruments[27,29–32,36,38–40]. 60% of the results were in accordance with the hypotheses (3 out of 5).

Criterion validity

Criterion validity could not be evaluated, because no studies compared the KOOS-PS or HOOS-PS with summed full-length HOOS or KOOS function and sports subscales.

Table 2.6. Hypotheses testing for construct validity and responsiveness.

	Construct validity		Responsiveness
	HOOS-PS	KOOS-PS	KOOS-PS
1	*accepted* 100% of the correlations with instruments measuring physical function had a correlation of ≥0.50	*accepted* 92% of the correlations with instruments measuring physical function had a correlation of ≥ 0.50	*rejected* 33% of the correlations with instruments measuring physical function had a correlation of ≥0.50
2	*rejected* 14% of the correlations with instruments measuring pain, stiffness or a combination of physical function and pain were 0.30-0.50	*rejected* 8% of the correlations with instruments measuring pain, stiffness or a combination of physical function and pain were 0.30-0.50	*rejected* 33% of the correlations with instruments measuring pain or a combination of physical function and pain were 0.30-0.50
3	*rejected* 60% of the correlations with instruments measuring unrelated constructs like mental health or social functioning were <0.30	*rejected* 27% of the correlations with instruments measuring unrelated constructs like mental health or social functioning were <0.30	*accepted* 100% of the correlations with instruments measuring unrelated constructs like mental health or self-efficacy were <0.30
4	*accepted* The mean correlation with instruments measuring similar constructs differed 0.156 from the mean correlation with instruments measuring related but dissimilar constructs	*accepted* The mean correlation with instruments measuring similar constructs differed 0.13 from the mean correlation with instruments measuring related but dissimilar constructs	*rejected* Mean correlation of instruments measuring similar constructs differed 0.06 from the mean correlation of instruments measuring related but dissimilar constructs
5	*accepted* Mean correlation with instruments measuring related but dissimilar constructs differed 0.34 from instruments measuring unrelated constructs	*accepted* Mean correlation with instruments measuring related but dissimilar constructs differed 0.27 from instruments measuring unrelated constructs.	*accepted* Mean correlation with instruments measuring related but dissimilar constructs differed 0.32 from instruments measuring unrelated constructs.

Responsiveness

There was very low quality evidence for indeterminate responsiveness of the HOOS-PS. All studies used the standardized response mean (SRM) to evaluate the responsiveness of the HOOS-PS[27,28,30,31,36,38,41]. The SRM can be used as an indirect measure when the expected change in health status is known, however it is not the preferred method. Since the expected change in health status on the construct of interest is not known, the standardized response mean cannot be used for evaluating responsiveness of the HOOS.PS[15].

There was high quality evidence for insufficient responsiveness of the KOOS-PS in patients with knee osteoarthritis receiving conservative treatment. Two studies with a very good methodology were pooled[27,41]. 40% of the results were in accordance with the hypotheses (Table 2.6). Both included studies assessed the correlations between

changes of the KOOS-PS with comparator measurement instruments, with predefined hypotheses in patients with knee osteoarthritis receiving conservative treatment. 13 correlations of changes in the KOOS-PS with comparator measurement instruments were found. All other studies evaluated responsiveness had an inadequate methodology and were excluded[27,28,30,31,33,36,38,41] because of using an inappropriate measure of responsiveness.

Interpretability

Table 2.7 presents the summary of the interpretability. It shows the weighted average score and standard deviation on the HOOS-PS and KOOS-PS in patients with osteoarthritis, after total joint replacement or conservative treatment. Furthermore, the minimally important change, the smallest detectable change and the patient acceptable symptom state are presented. There were no floor or ceiling effects.

Table 2.7. Interpretability: average scores, floor and ceiling effects, MIC/PASS/SDC values for HOOS-PS and KOOS-PS.

	Weighted average score (SD; n)			Anchor based values				
	Osteo-arthritis	Post THR/ TKR	Post conservative treatment	MIC	PASS	SDC	Floor/ ceiling effects	Missing items
HOOS-PS	56.7 (20; 4084) (21,28–30,36–38)	20.1 (19; 2949)(21,28,37)	41.3 (16.2; 20) (36)	23 (CI 19-30) (37)	88 (CI 87-88) (37)	NR	None (34,36)	0-3% (34,36,37)
KOOS-PS	52.9 (17.6; 4651) (27–32,36,38,40,41)	34 (16.6; 2289) (21,28,33)	38.4 (18.4; 257)(27,31,36)	2.2 (SD 17.5) and 12.0 (27,35)	NR	16 and 28.3 (27,35)	<0.01% ceiling, <2.4% floor; none; ceiling 5%, floor 0.4%; (33,36,40)	0%; 0%; 7-11.7%, squatting and kneeling items missing 4-6% post TKR (31,36)

Abbreviations: SD= standard deviation; n= number of patients; THR/TKR= total hip replacement, total knee replacement; NR= not reported; PASS= patient acceptable symptom state; MIC= minimally important change; CI= confidence interval; SDC= smallest detectable change.

Feasibility

Table 2.8 shows an overview of the feasibility. The authors described the application of the measurement instruments as easy to use, short, and free of charge and copyright.

Table 2.8. Feasibility of the HOOS-PS and KOOS-PS, table based on the COSMIN manual and the guideline for selecting PROMs for Core Outcome Sets.[16,47]

Feasibility aspects	HOOS-PS	KOOS-PS
Patients comprehensibility	Not evaluated, assumed to be good	Not evaluated, assumed to be good
Clinician's comprehensibility	Good	Good
Type and ease of administration	Self-administered, easy to use	Self-administered, easy to use
Length of the instrument	Short, 5 items	Short, 7 items
Completion time	Not registered, assumed to be maximal 3 minutes	Not registered, assumed to be maximal 3 minutes
Patient's required mental and physical ability level	Usage >13 years, mentally competent, all patients with hip complaints	Usage >13 years, mentally competent, all patients with knee complaints
Ease of standardization	No data available	No data available
Ease of score calculation	Easy	Easy
Copyright	Permission not required to use the HOOS-PS	Permission not required to use the KOOS-PS
Cost of an instrument	Free of charge	Free of charge
Required equipment	Paper or online	Paper or online
Availability in different settings	Self-administered. No interview or phone formats are available	Self-administered. No interview or phone formats are available
Regulatory agency's requirement for approval	Not known	Not known

Discussion

The present study determined the current evidence on the measurement properties of the HOOS-PS and KOOS-PS. The most important finding was the observed lack of several components of the validity of the HOOS-PS and KOOS-PS, such as content validity and construct validity. This implies that the scores on the HOOS-PS and KOOS-PS may inadequately reflect physical functioning in patients with hip or knee complaints. Furthermore, there is evidence for insufficient construct validity and responsiveness in patients with knee osteoarthritis receiving conservative treatment.

All outcome scores and data on measurement properties must be interpreted with caution because the content validity of both outcome measurement instruments was inconclusive. This means that it is unclear if the HOOS-PS and KOOS-PS adequately reflect physical functioning. This can be explained by concerns regarding the

relevance and the comprehensiveness of the items of the questionnaires. The unclear content validity can possibly interfere with outcomes on all other measurement properties and should be taken into account when evaluating and interpreting them.

An implication of the problematic validity is the assumption that the HOOS-PS is a reliable outcome measurement instrument, however it cannot be confirmed that the HOOS-PS is reliably measuring the construct physical functioning solely and comprehensively. The found correlations between the HOOS-PS and KOOS-PS with instruments measuring different constructs like pain and stiffness were higher than hypothesized; indicating that they may be measuring a broader construct than just physical functioning. For example, constructs of physical functioning, pain and stiffness may theoretically be distinguishable; however, patients may respond globally. Regarding the HOOS-PS and KOOS-PS, it is possible that the difficulty during activity experienced by patients is influenced by the degree of pain, physical functioning or stiffness. The inability to distinguish between pain and physical functioning was demonstrated for other outcome measurement instruments[42,43]. Concerning the KOOS-PS, of the eight measurement properties, most were rated as indeterminate or inconsistent and only the reliability was sufficient. This could be due to the inconsistent evidence on content validity.

This is the first review evaluating the responsiveness of the KOOS-PS with studies using adequate methodology, whereas earlier reviews only considered inadequate measures for responsiveness as SRM and effect size[10,13]. The responsiveness was rated as insufficient, indicating that the KOOS-PS is limited in detecting improvement in physical function over time in patients with knee osteoarthritis receiving conservative treatment and thus is probably not the most suitable instrument for measuring outcomes in this population. A previous review evaluating the structural validity concluded that the majority of evidence suggested a unidimensional structure of the KOOS-PS[13], however we were unable to find evidence to support this. Evidence for the structural validity of the KOOS-PS was rated as inconsistent and because of this, we rated the evidence for internal consistency as indeterminate.

A strength of this systematic review is that this review was conducted according to the COSMIN guideline for systematic reviews[15]. The research question was answered extensively and completely. As the HOOS-PS and KOOS-PS are widely used measurement instruments, we could evaluate more evidence on the measurement properties than previous reviews. We evaluated the PROM development and content validity in a systematic and qualitative manner and used not only the PROM development studies of

the short forms but also the item development studies of the full-length HOOS, KOOS and WOMAC to obtain information.

The limitations of this study were in particular caused by the inadequate methodology of the included studies. Many measurement properties were evaluated with an inadequate methodology making them unusable to include in this review and more importantly, the inadequate methodology can lead to incorrect conclusions in the studies in question. For instance, with regard to reliability several studies included less than 50 patients. None of the studies used an adequate method to evaluate cross-cultural validity, despite several translations intended to[34], therefore, cross-cultural validity could not be evaluated. Of all included articles, six articles extracted the HOOS-PS and/or KOOS-PS scores out of the full-length HOOS or KOOS. This may have influenced the outcome or missing data. While internal consistency, reliability, construct validity and responsiveness were assessed frequently, the properties content validity, structural validity, cross-cultural validity and criterion validity were not.

Further research should evaluate the content validity of the HOOS-PS and KOOS-PS in more detail and focus on improving the relevance and comprehensiveness of the items to better measure the construct of physical functioning. Alternatively, other PROMs can be explored, for example the promising computer adaptive testing with Patient-Reported Outcomes Measurement Information System (PROMIS)[44]. Although not encouraged, when developing new measurement instruments, it is recommended to use a theoretical model (for example the International Classification of Functioning (ICF[45])) to map the construct to a model. Furthermore, future studies assessing measurement properties of PROMs, are recommended to use the COSMIN design checklist as a guide for achieving adequate methodology[17].

This review and meta-analysis shows that the widespread use in clinical practice of the HOOS-PS and KOOS-PS is not scientifically supported. Although we found evidence for sufficient reliability, the inconsistent evidence on content validity and the insufficient construct validity and responsiveness has implications on all these properties. It may be that the HOOS-PS and KOOS-PS may not measure physical functioning solely and comprehensively. Concluding, scores on the HOOS-PS and KOOS-PS may inadequately reflect physical functioning in patients undergoing total hip and total knee arthroplasty and in patients with hip or knee complaints. Possible consequences of continuing using these questionnaires, are incorrect interpretation of the outcome scores of the individual patients and average outcome scores of healthcare providers with possible patient- and hospital related consequences. Using the HOOS-PS or KOOS-PS as outcome

measurement instruments for comparing outcomes, measuring improvements or benchmarking in patients with hip or knee complaints or undergoing arthroplasty should only be done with great caution.

Contributions

Conception and design of the study: CB, NW, SP, RV, RO; Analysis and interpretation of the data: CB, NW, SP, RV, RO; Drafting of the article: CB, NW; Critical revision of the article for important intellectual content: CB, NW, SP, RV, RO, Dr. C.B. Terwee; Final approval of the article: CB, NW, SP, RV, RO; statistical expertise: CB, NW, SP, RO, Prof. Dr. HCW de Vet; Collection and assembly of data: CB, NW, SP.

Christel Braaksma (christelbraaksma@hotmail.com) and Nienke Wolterbeek (orthopedie-research@antoniusziekenhuis.nl) take responsibility for the integrity of the work as a whole, from inception to finished article.

Conflict of interest

None of the authors had conflicts of interests.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The authors did not have any other financial interests what could lead to a conflict of interest regarding to this study.

Acknowledgments

The authors want to thank Dr. associate Prof. C.B. Terwee, associate professor in Measurement, department of epidemiology and biostatistics at Amsterdam University Medical Centre, location VUmc for her critically appraising and improving this manuscript. Furthermore, the authors want to thank Prof. Dr. H.C.W. de Vet, professor in clinimetrics at the VU university medical centre Amsterdam for her feedback, especially checking the methodology of the meta-analysis.

References

1. Black N. Patient reported outcome measures could help transform healthcare. BMJ. 2013 Jan 28;346:f167.
2. Davis AM, Perruccio A V., 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. Osteoarthr Cartil. 2008 May;16(5):551–9.
3. Perruccio A V., Stefan Lohmander L, Canizares M, Tennant A, Hawker GA, Conaghan PG, et al. The development of a short measure of physical function for knee OA KOOS-Physical Function Shortform (KOOS-PS) - an OARSI/OMERACT initiative. Osteoarthr Cartil. 2008 May;16(5):542–50.
4. Bellamy N, Buchanan WW. A preliminary evaluation of the dimensionality and clinical importance of pain and disability in osteoarthritis of the hip and knee. Clin Rheumatol. 1986 Jun; 5(2):231-41.
5. Klässbo 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; 32(1):46-51.
6. Roos EM, Roos HP, Lohmander LS, Ekdahl C, Beynnon BD. Knee Injury and Osteoarthritis Outcome Score (KOOS) - Development of a self-administered outcome measure. J Orthop Sports Phys Ther. 1998 Aug; 28(2): 88-96.
7. Dahlberg LE. ICHOM Standard Set for monitoring knee and hip osteoarthritis. Osteoarthr Cartil. 2016 Apr;24:S436–7.
8. Aveline C, Roux A Le, Hetet H Le, Gautier JF, Vautier P, Cognet F, et al. Pain and recovery after total knee arthroplasty: A 12-month follow-up after a prospective randomized study evaluating nefopam and ketamine for early rehabilitation. Clin J Pain. 2014 Sep;30(9):749–54.
9. Gossec L, Paternotte S, Maillefert JF, Combescure C, Conaghan PG, Davis AM, et al. The role of pain and functional impairment in the decision to recommend total joint replacement in hip and knee osteoarthritis: An international cross-sectional study of 1909 patients. Report of the OARSI-OMERACT Task Force on total joint replacement. Osteoarthr Cartil. 2011 Feb;19(2):147–54.
10. Gagnier JJ, Mullins M, Huang H, Marinac-Dabic D, Ghambaryan A, Eloff B, et al. A Systematic Review of Measurement Properties of Patient-Reported Outcome Measures Used in Patients Undergoing Total Knee Arthroplasty. Journal of Arthroplasty. 2017 May;32(5):1688-1697.e7.
11. Gagnier JJ, Huang H, Mullins M, Marinac-Dabić D, Ghambaryan A, Eloff B, et al. Measurement properties of patient-reported outcome measures used in patients undergoing total hip arthroplasty: A systematic review. JBJS Rev. 2018 Jan;6(1):e2.
12. Peer MA, Lane J. The knee injury and osteoarthritis outcome score (KOOS): A review of its psychometric properties in people undergoing total knee arthroplasty. J Orthop Sports Phys Ther. 2013 Jan;43(1):20–8.
13. Collins NJ, Prinsen CAC, Christensen R, Bartels EM, Terwee CB, Roos EM. Knee Injury and Osteoarthritis Outcome Score (KOOS): systematic review and meta-analysis of measurement properties. Osteoarthr Cartil. 2016 Aug;24(8):1317–29.
14. Shamseer L, Moher D, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. Preferred reporting items for systematic review and meta-analysis protocols (prisma-p) 2015: Elaboration and explanation. BMJ 2015 Jan 2;350:g7647.
15. Prinsen CAC, Mokkink LB, Bouter LM, Alonso J, Patrick DL, de Vet HCW, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res. 2018 May; 27(5):1147-57.
16. Mokkink LB, Prinsen CAC, Patrick DL, Alonso J, Bouter LM, de Vet HCW, et al. COSMIN methodology for systematic reviews of Patient-Reported Outcome Measures (PROMs) user manual [Internet]. 2018 [cited 2020 Feb 19]. Available from: https://www.cosmin.nl/wp-content/uploads/COSMIN-syst-review-for-PROMs-manual_version-1_feb-2018.pdf
17. Mokkink LB, de Vet HCW, Prinsen CAC, Patrick DL, Alonso J, Bouter LM, et al. COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. Qual Life Res. 2018 May;27(5):1171-9.
18. Terwee CB, Mokkink LB, Knol DL, Ostelo RWJG, Bouter LM, De Vet HCW. Rating the methodological quality in systematic reviews of studies on measurement properties: A scoring system for the COSMIN checklist. Quality of Life Research. 2012 May;21(4):651-7.

19. Feldt LS, Charter RA. Averaging internal consistency reliability coefficients. Educ Psychol Meas. 2006 April;66(2):215-27
20. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986 Sep;7(2):177-88.
21. Wiering B, de Boer D, Delnoij D. Asking what matters: The relevance and use of patient-reported outcome measures that were developed without patient involvement. Health Expect. 2017 Dec;20(6):1330–41.
22. 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. Osteoarthr Cartil. 2007 Jan;15(1):104–9.
23. de Groot IB, Favejee MM, Reijman M, Verhaar JAN, Terwee CB. The dutch version of the knee injury and osteoarthritis outcome score: A validation study. Health Qual Life Outcomes. 2008 Feb;26:6-16.
24. Roos EM, Roos HP, Ekdahl C, Lohmander LS. Knee injury and Osteoarthritis Outcome Score (KOOS) - validation of a Swedish version. Scand J Med Sci Sports. 2007;8(6):439–48.
25. 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;63(7):737-45.
26. Franchignoni F, Salaffi F, Giordano A, Carotti M, Ciapetti A, Ottonello M, et al. Rasch analysis of the 22 knee injury and osteoarthritis outcome score-physical function items in italian patients with knee osteoarthritis. Arch Phys Med Rehabil. 2013 Mar;94(3):480–7.
27. Harris KK, Dawson J, Jones LD, Beard DJ, Price AJ. Extending the use of PROMs in the NHS-using the Oxford Knee Score in patients undergoing non-operative management for knee osteoarthritis: A validation study. BMJ Open. 2013 Aug;3(8):e003365.
28. Davis AM, Perruccio A V, Canizares M, Hawker GA, Roos EM, Maillefert J-F, et al. Comparative, validity and responsiveness of the HOOS-PS and KOOS-PS to the WOMAC physical function subscale in total joint replacement for osteoarthritis. Osteoarthr Cartil. 2009 Jul;17(7):843–7.
29. Mehta SP, Sankar A, Venkataramanan V, Lohmander LS, Katz JN, Hawker GA, et al. Cross-cultural validation of the ICOAP and physical function short forms of the HOOS and KOOS in a multi-country study of patients with hip and knee osteoarthritis. Osteoarthr Cartil. 2016 Dec;24(12):2077–81.
30. Ruyssen-Witrand A, Fernandez-Lopez CJ, Gossec L, Anract P, Courpied JP, Dougados M, et al. Psychometric properties of the OARSI/OMERACT osteoarthritis pain and functional impairment scales: ICOAP, KOOS-PS and HOOS-PS. Clin Exp Rheumatol. 2011;29(2):231–7.
31. Gonçalves RS, Cabri J, Pinheiro JP, Ferreira PL, Gil J. Reliability, validity and responsiveness of the Portuguese version of the Knee injury and Osteoarthritis Outcome Score - Physical Function Short-form (KOOS-PS). Osteoarthr Cartil. 2010 Mar;18(3):372–6.
32. Gul ED, Yilmaz O, Bodur H. Reliability and validity of the Turkish version of the knee injury and osteoarthritis outcome score-physical function short-form (KOOS-PS). J Back Musculoskelet Rehabil. 2013;26(4):461–6.
33. Gandek B, Roos EM, Franklin PD, Ware JE. Item selection for 12-item short forms of the Knee injury and Osteoarthritis Outcome Score (KOOS-12) and Hip disability and Osteoarthritis Outcome Score (HOOS-12). Osteoarthr Cartil. 2019;27(5):746–53.
34. Yilmaz O, Gul ED, Bodur H. Cross-cultural adaptation and validation of the Turkish version of the Hip disability and Osteoarthritis Outcome Score-Physical function Short-form (HOOS-PS). Rheumatol Int. 2014 Jan;34(1):43–9.
35. Singh JA, Luo R, Landon GC, Suarez-Almazor M. Reliability and clinically important improvement thresholds for osteoarthritis pain and function scales: A multicenter study. J Rheumatol. 2014 Mar;41(3):509–15.
36. Ornetti P, Perruccio A V, Roos EM, Lohmander LS, Davis AM, Maillefert JF, et al. Psychometric properties of the French translation of the reduced KOOS and HOOS (KOOS-PS and HOOS-PS). Osteoarthr Cartil. 2009 Dec;17(12):1604–8.
37. 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. Acta Orthop. 2014 Feb;85(1):39–48.
38. Bond M, Davis A, Lohmander S, Hawker G. Responsiveness of the OARSI-OMERACT osteoarthritis pain and function measures. Osteoarthr Cartil. 2012 Jun;20(6):541–7.

39. Davis AM, Badley EM, Hogg-Johnson S, Ibrahim S, Perruccio AV, Wong R, et al. Understanding early recovery following primary total hip and knee replacement. Arthritis Rheum. 2009;60:1938.
40. Stratford PW, Kennedy DM. A Comparison Study of KOOS-PS and KOOS Function and Sport Scores. Phys Ther. 2014 Nov;94(11):1614–21.
41. Mahler E, Cuperus N, Bijlsma J, Vliet Vlieland T, van den Hoogen F, den Broeder AA, et al. Responsiveness of four patient-reported outcome measures to assess physical function in patients with knee osteoarthritis. Scand J Rheumatol. 2016 Nov;45(6):518–27.
42. Terwee CB, van der Slikke RMA, van Lummel RC, Benink RJ, Meijers WGH, de Vet HCW. Self-reported physical functioning was more influenced by pain than performance-based physical functioning in knee-osteoarthritis patients. J Clin Epidemiol. 2006 Jul;59(7):724-31.
43. Stratford P, Kennedy D, Clarke H. Confounding pain and function: the WOMAC's failure to accurately predict lower extremity function. Arthroplast Today. 2018 Oct;4(4):488-92.
44. Rose M, Bjorner JB, Gandek B, Bruce B, Fries JF, Ware JE. The PROMIS Physical Function item bank was calibrated to a standardized metric and shown to improve measurement efficiency. J Clin Epidemiol. 2014 May;67(5):516-26.
45. World Health Organization. Towards a Common Language for Functioning , Disability and Health ICF. International Classification. Phys Ther 2002 May;86(5):726-34.

Supplemental material

Search strategy:

PubMed

(hip osteoarthritis outcome[tiab] OR hoos[tiab] OR knee osteoarthritis outcome[tiab] OR koos[tiab] OR ((knee injury[tiab] OR hip disability) AND osteoarthritis outcome score[tiab])) AND (ps[tiab] OR short form*[tiab] OR physical function*[tiab])

Cochrane library

("hip osteoarthritis outcome" OR hoos OR "knee osteoarthritis outcome" OR koos OR (("hip disability" OR "knee injury") NEAR/2 "osteoarthritis outcome score")):ab,ti AND (ps OR "short form*" OR "physical function*"):ab,ti

Embase (via Embase.com)

('hip osteoarthritis outcome' OR hoos OR 'knee osteoarthritis outcome' OR koos OR ((‘hip disability’ OR ‘knee injury’) NEAR/2 ‘osteoarthritis outcome score’)):ab,ti AND (ps OR 'short form*' OR 'physical function*'):ab,ti NOT 'conference abstract'/it

Cinahl Plus (via EBSCOhost)

(TI ('hip osteoarthritis outcome' OR hoos OR 'knee osteoarthritis outcome' OR koos OR ((‘knee injury’ OR ‘hip disability’) AND ‘osteoarthritis outcome score’)) OR AB ('hip osteoarthritis outcome' OR hoos OR 'knee osteoarthritis outcome' OR koos OR ((‘knee injury’ OR ‘hip disability’) AND ‘osteoarthritis outcome score’))) AND (TI (ps OR 'short form*' OR 'physical function*') OR AB (ps OR 'short form*' OR 'physical function*'))

PsycInfo (via Ovid)

('hip osteoarthritis outcome' or hoos or 'knee osteoarthritis outcome' or koos or (('hip disability' or 'knee injury') adj2 'osteoarthritis outcome score')).ti,ab.

Updated criteria for good measurement properties, obtained from Prinsen et al. 2018.[15]

Measurement property	Rating[1]	Criteria
Structural validity	+	**CTT:** CFA: CFI or TLI or comparable measure >0.95 OR RMSEA <0.06 OR SRMR <0.08[2] **IRT/Rasch**: No violation of unidimensionality[3]: CFI or TLI or comparable measure >0.95 OR RMSEA <0.06 OR SRMR <0.08 *AND* no violation of local independence: residual correlations among the items after controlling for the dominant factor <0.20 OR Q3's <0.37 *AND* no violation of monotonicity: adequate looking graphs OR item scalability >0.30 *AND* adequate model fit: IRT: χ^2 >0.01 Rasch: infit and outfit mean squares ≥0.5 and ≤1.5 OR Z-standardized values >-2 and <2
	?	CTT: Not all information for '+' reported IRT/Rasch: Model fit not reported
	–	Criteria for '+' not met
Internal consistency	+	At least low evidence[4] for sufficient structural validity[5] AND Cronbach's alpha(s) ≥0.70 for each unidimensional scale or subscale[6]
	?	Criteria for "At least low evidence[4] for sufficient structural validity[5]" not met
	–	At least low evidence[4] for sufficient structural validity[5] AND Cronbach's alpha(s) <0.70 for each unidimensional scale or subscale[6]
Reliability	+	ICC or weighted Kappa ≥0.70
	?	ICC or weighted Kappa not reported
	–	ICC or weighted Kappa <0.70
Measurement error	+	SDC or LoA <MIC[5]
	?	MIC not defined
	–	SDC or LoA >MIC[5]
Hypotheses testing for construct validity	+	The result is in accordance with the hypothesis[7]
	?	No hypothesis defined (by the review team)
	–	The result is not in accordance with the hypothesis[7]

Measurement property	Rating[1]	Criteria
Cross-cultural validity\measurement invariance	+	No important differences found between group factors (such as age, gender, language) in multiple group factor analysis OR no important DIF for group factors (McFadden's R^2 <0.02)
	?	No multiple group factor analysis OR DIF analysis performed
	–	Important differences between group factors OR DIF was found
Criterion validity	+	Correlation with gold standard ≥0.70 OR AUC ≥0.70
	?	Not all information for '+' reported
	–	Correlation with gold standard <0.70 OR AUC <0.70
Responsiveness	+	The result is in accordance with the hypothesis[7] OR AUC ≥0.70
	?	No hypothesis defined (by the review team)
	–	The result is not in accordance with the hypothesis[7] OR AUC <0.70

AUC = area under the curve, CFA = confirmatory factor analysis, CFI = comparative fit index, CTT = classical test theory, DIF = differential item functioning, ICC = intraclass correlation coefficient, IRT = item response theory, LoA = limits of agreement, MIC = minimal important change, RMSEA: Root Mean Square Error of Approximation, SEM = Standard Error of Measurement, SDC = smallest detectable change, SRMR: Standardized Root Mean Residuals, TLI = Tucker-Lewis index

[1] "+" = sufficient, " –" = insufficient, "?" = indeterminate
[2] To rate the quality of the summary score, the factor structures should be equal across studies
[3] Unidimensionality refers to a factor analysis per subscale, while structural validity refers to a factor analysis of a (multidimensional) patient-reported outcome measure
[4] As defined by grading the evidence according to the GRADE approach
[5] This evidence may come from different studies
[6] The criteria 'Cronbach alpha <0.95' was deleted, as this is relevant in the development phase of a PROM and not when evaluating an existing PROM.
[7] The results of all studies should be taken together and it should then be decided if 75% of the results are in accordance with the hypotheses

CHAPTER 3

The Hip Disability and Osteoarthritis Outcome Score-Physical Function Shortform Does Not Adequately Represent Physical Functioning in Patients Undergoing Total Hip Arthroplasty

Value in Health 2022 Nov; 25(11):1894-1901

Abstract

Objectives

A frequently used Patient Reported Outcome Measure for assessing physical functioning in patients with hip problems is the 5-item Hip disability and Osteoarthritis Outcome Score-Physical function Shortform (HOOS-PS). However, its content validity (whether this instrument adequately reflects the construct of physical functioning) is unknown. The aim of this study was to assess the content validity of the HOOS-PS.

Methods

A quantitative and qualitative research approach was used. Physical functioning was defined as the ability to carry out activities that require physical actions, ranging from self-care to more complex activities that require a combination of skills, often within a social context. Patients (n=51) and experts (n=25) completed questionnaires regarding the relevance, comprehensiveness and comprehensibility of the HOOS-PS. Semi-structured interviews (n=5) explored issues identified in the quantitative data in more depth. Thematic content analysis was carried out using a coding frame.

Results

One of the five items was considered relevant to measure physical functioning. Comprehensiveness was considered insufficient by both patients and experts. Furthermore, comprehensibility was considered inadequate. Several items were found ambiguous or double-barrelled. Regarding interpretability, floor or ceiling effects were not found.

Conclusions

This study showed concerns about the content validity of the HOOS-PS: the majority of the items are considered not relevant, the HOOS-PS is considered not comprehensive and several items are considered not comprehensible. These findings challenge the applicability of the HOOS-PS in clinical practice, research, VBHC and benchmarking.

Introduction

Patient reported outcome measures (PROMs) are frequently used to evaluate health outcomes. Therefore, it is important that these instruments truly measure what they intend to measure. This is called validity. Content validity is an aspect of validity and is defined as the degree to which the content of a PROM is 'an adequate reflection of the construct to be measured'[1]. It is considered the most important measurement property of a PROM. Content validity refers to relevance, comprehensiveness and comprehensibility[1]. Relevance is defined as the degree to which the content is considered applicable for measuring physical function. Comprehensiveness reflects the degree to which all key aspects of the construct are covered in the measurement instrument. Comprehensibility is defined as the degree to which the items, response options and instructions are understood by patients as intended[1].

A frequently used PROM for assessing physical functioning in patients with hip problems is the Hip disability and Osteoarthritis Outcome Score-Physical function Shortform (HOOS-PS). The International Consortium for Health Outcomes Measurement (ICHOM) recommend the use of the HOOS-PS in clinical practice and several joint registries included the HOOS-PS as a measure of physical functioning in their standard outcome sets for evaluating health outcomes in patients with hip osteoarthritis[2,3]. The HOOS-PS was developed by Davis et al.[4] using Rasch analysis. A recent systematic review of the measurement properties of the HOOS-PS showed an inconsistent evidence for content validity, because not all aspects of content validity were determined and the quality of evidence was low[5].

It is essential to understand the adequacy in which the HOOS-PS reflects physical functioning before the measurement instrument is used in clinical practice, research, Value Based Healthcare (VBHC), and benchmarking. It is therefore necessary to determine its content validity to assess if the HOOS-PS can reliably measure physical functioning in patients with hip problems. The aim of this study was to assess the content validity of the HOOS-PS in patients with hip problems. This paper explores the views of patients and experts regarding the relevance, comprehensiveness and comprehensibility of the HOOS-PS.

Methods

This combined quantitative and qualitative prospective study was conducted in accordance with the COnsolidated criteria for REporting Qualitative research (COREQ) and Standards for Reporting Qualitative Research (SRQR)[6,7]. Patients and physicians completed study-specific questionnaires, reflecting on the content validity of the HOOS-PS. Complementary semi-structured interviews were conducted for more in-depth understanding and interpretation for the acquired data. The study was approved by the Institutional Review Board of the St. Antonius Hospital.

HOOS-PS

The HOOS-PS[4] was developed in commission by the Osteoarthritis Research Society International (OARSI) to measure the domain physical functioning in hip osteoarthritis, by the Outcome Measures in Rheumatology Clinical Trials (OMERACT) working group[8]. The HOOS-PS contains five items, adapted in unchanged form from the Western Ontario McMaster Universities Osteoarthritis Index (WOMAC, 3 items)[9] and the Hip disability and Osteoarthritis Outcome Score (HOOS, 2 items)[10]. For each item, patients have to indicate the degree of difficulty experienced due to their hip problem in the last week. The translation and basic validation of the Dutch HOOS-PS was performed by De Groot et al.[11].

The construct of physical functioning

Since the developers of the HOOS-PS did not base the targeted construct on a theoretical model, there is a need to define the construct 'physical functioning'. Neither the World Health Organization nor the International Classification of Functioning defined the construct[12]. A study revising domain definitions, defined physical functioning as "the ability to carry out activities that require physical actions, ranging from self-care (activities of daily living) to more complex activities that require a combination of skills, often within a social context"[13]. We adopted this definition to assess content validity.

Eligibility

Potential participants were patients identified and recruited at a district general hospital (St. Antonius Hospital, Utrecht), including patients scheduled for a total hip arthroplasty (THA) who were present at one of the weekly patient information meetings. Over a period of three months, all eligible patients were approached to participate in a questionnaire study, or in an interview. Inclusion criteria for patients were age 18 years or older and having hip problems for which a THA was scheduled. Exclusion criteria were

insufficient understanding of the Dutch language and cognitive impairment. As a result, a demographic diversity regarding age, gender and background was expected. Patients or experts did not receive any payment for taking part. At least fifty participants were included to meet the standards of the COSMIN group for a good content validity study[14]. Furthermore, twenty-five experts who were involved in the care of patients undergoing THA on a daily basis of two medium sized teaching hospitals in the Netherlands were recruited (St. Antonius Hospital, Utrecht and OLVG, Amsterdam).

Questionnaires and interviews

The COSMIN methodology for assessing the content validity of PROMs - User Manual[15] was used to create a coding frame (Figure 3.1) and the study-specific questionnaire. The verification if the HOOS-PS reflected physical functioning was assessed per domain of content validity. The relevance and comprehensiveness were tested by both questionnaires (patients and experts) and interviews. The comprehensibility was analyzed during the interviews.

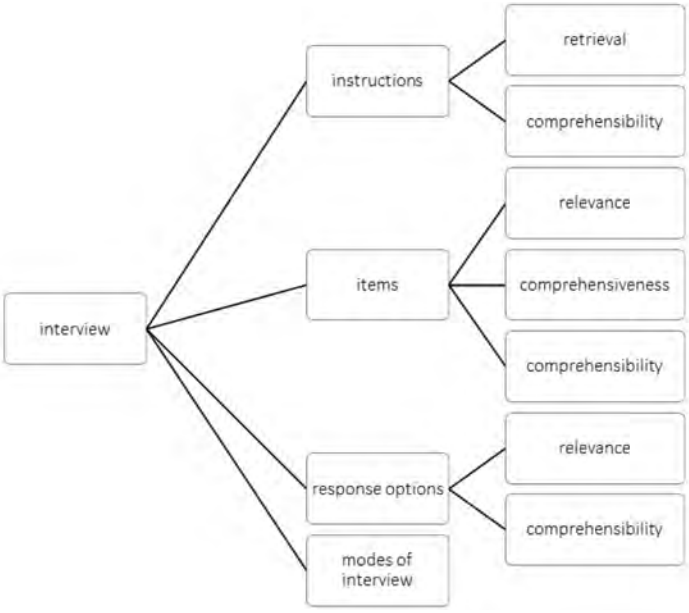

Figure 3.1. Coding frame consisting of the applicable components of content validity (relevance, comprehensiveness, comprehensibility) per part of the HOOS-PS (instruction, items, response options) for the merged data (questionnaires & interviews).

At baseline, patients filled out the questionnaire in which they were asked to score each of the five items as relevant or irrelevant in relation to their hip problems (binary

response option). Furthermore, they were instructed to reflect on the content validity in several open-ended questions. For instance, it was verified if the key aspects of physical functioning were covered and if any activities were missing (*if yes, explain*). Since physical functioning is expected to change due to the THA, patients received the same paper questionnaires three months after THA. The questionnaire was sent by mail with a stamped addressed envelope for reply. Non-responders were approached by telephone. After analyzing the concepts of interest of the questionnaires, the coding frame was adjusted.

Cognitive semi-structured interviews were conducted to get in-depth data relating to questionnaire completion regarding the relevance, comprehensiveness, and comprehensibility of the HOOS-PS. An interview-guide was developed for the semi-structured interview based on the Three-Step Test-Interview method[16] and the guideline of Schildmann *et al* (2016)[17]. The coding frame was used as structure for the interview. Patients were recruited to have demographic diversity regarding age, gender and background. None of the patients refused to participate or dropped out. A study information form was provided and written informed consent was obtained from all participants. The interviews took place at the outpatient clinic, with the patient sometimes accompanied by a relative. The interviews were conducted by a clinician-researcher (CB), experienced in qualitative research[18] and not involved in the clinical care of the patients. During the interview, patients were asked to reflect on the relevance of each item of the HOOS-PS. Comprehensiveness was assessed by asking if they felt that this set of questions covered all aspects of the physical functioning of their hip. Furthermore, patients were asked whether the questionnaire missed an activity that is more important for their hip problems or if there were items they would remove from the questionnaire. Finally, comprehensibility of the instructions, items and response options was tested. This was assessed using both 'think aloud' technique and verbal probing techniques to observe difficulties in completing the HOOS-PS[16]. The interviews were recorded and transcribed verbatim. The interviews were performed until code saturation was reached.

Experts were requested to review the HOOS-PS and then complete an online questionnaire, rating the content validity of the measure. First, they had to rate each question of the HOOS-PS as relevant or irrelevant. Furthermore, experts had to specify if the question was relevant for all THA patients, if the question was discriminative for the level of functioning and if there is an expected change following arthroplasty (Figure 2a and 2b). For each question, they were asked to explain their answer. Additionally, they were asked about comprehensiveness of the measurement instrument.

Rating

The rating of the content validity of the HOOS-PS was conducted according to the criteria for good content validity of the COSMIN group (Table 3.1)[14]. Each item was considered relevant if at least 85% of the patients thought that item was relevant. If at least 85% of the items were concerned relevant for measuring physical functioning by the target population and experts, the total set of items was considered relevant (Table 3.1). The PROM was considered comprehensive when patients and experts found that no key concepts were missing. Experts had to find the PROM comprehensive for measuring physical functioning, for this specific population and in this context of use (measuring physical functioning in patient undergoing total hip arthroplasty). Comprehensibility was found adequate if at least 85% of the items and response options were found comprehensible by patients (Table 3.1).

Table 3.1. Criteria for evaluating the content validity, according to the COSMIN criteria and ratings of the content validity of the HOOS-PS in this study[14,15].

Criteria	Guidance for giving a sufficient (+) rating	Rating Sufficient (+), insufficient (-), inconsistent (±)
Relevance*		
1. Are the included items relevant for the construct of interest?	Professionals considered ≥85% of the items relevant for the construct of interest	-
2. Are the included items relevant for the target population of interest?	Patients rated ≥85% of the items relevant	-
3. Are the included items relevant for the context of use of interest?	Professionals considered ≥85% relevant for the context of use	-
4. Are the response options appropriate?	Patients rated ≥85% of the response options relevant	+
5. Is the recall period appropriate?	Patients rated the recall period appropriate	+
RELEVANCE RATING		±
* Each item was considered relevant if at least 85% of the patients thought that item was relevant. If at least 85% of the items were concerned relevant for measuring physical functioning by the target population and experts, the total set of items was considered relevant.		
Comprehensiveness*		
6. Are all key concepts included?	Patients and professionals consider the PROM comprehensive for the construct, population and context of use	-
COMPREHENSIVENESS RATING		-
* The PROM was considered comprehensive when patients and experts found that no key concepts were missing. Experts had to find the PROM comprehensive for measuring physical functioning, for this specific population and in this context of use (measuring physical functioning in patient undergoing total hip arthroplasty).		

Table 3.1. (continued)

Comprehensibility*		
7. Are the PROM instructions understood by the population of interest as intended?	No important problems were found	+
8. Are the PROM items and response options understood by the population of interest as intended?	No important problems were found for ≥85% of the items and response options.	-
9. Are the PROM items appropriately worded?	Reviewers consider ≥85% appropriately worded	-
10. Do the response options match the question?	Reviewers consider ≥85 of the response options matching the questions	+
COMPREHENSIBILITY RATING		-

* The comprehensibility was found adequate if at least 85% of the items and response options were found comprehensible by patients.

Statistical analysis

Statistical analysis of the quantitative data was performed using the statistical program SPSS® (Statistical Package for the Social Sciences, Chicago, IL, version 24.0). All input of patients during the interview regarding general impression, the instructions, burden and difficulties per item were collected and analyzed using thematic analysis. The questionnaire and interview data were merged in a text file per part of the HOOS-PS, namely the instruction, the individual items and response options. The open-ended answers of the questionnaire were added in the text file to the merged data per part of the HOOS-PS. Afterwards, each transcribed interview was thematically analyzed (CB and NW) using the coding frame (Figure 3.1). The data were collected in a database on a secured disk in the hospital.

Results

The paper questionnaire was completed at baseline by 51 patients, 90.2% of these patients filled out the postoperative assessment (Table 3.2). Each interview was transcribed and analyzed for new codes. There appeared no new concepts in the last two interviews, indicating data saturation. Twenty-five experts filled out the online questionnaire, of whom 5 were physical therapists, 11 orthopaedic surgeons and 9 residents orthopaedic surgery. Table 3.1 shows the ratings on the content validity. The results are described per section of content validity.

Table 3.2. Patient characteristics.

	Questionnaires	Interview
Number of patients	51	5
Male gender	21 (41%)	1 (20%)
Completed follow-up	46 (90%)	Not applicable
Mean age in years (SD)	69.0 (9.8)	65.0 (14.0)
Mean score HOOS-PS pre-surgery (SD)	48.7 (18.6)	Not applicable
Mean score HOOS-PS post-surgery (SD)	23.5 (17.1)	Not applicable
Mean duration interviews	Not applicable	13 minutes

Relevance

Overall, the relevance of the items in the HOOS-PS was found to be insufficient. The relevance reported by experts and patients preoperative and postoperative is shown per item in Table 3.3, Figure 3.2a and 3.2b. Patients considered two of five items of the HOOS-PS relevant in evaluating physical functioning *(descending stairs, twisting/pivoting on loaded leg)*. The experts considered only one of five items relevant *(descending stairs)* and the other four were considered as not referring to the construct of interest or were not relevant for the target population.

Table 3.3. Relevance per item and comprehensiveness according to patients and experts.

Item '*Indicate the degree of difficulty you have experienced last week due to your hip problem with…*'	Patients pre-op (n=51)	Patients post-op (n=46)	Experts (n=25)
1. descending stairs	83.7%	87.5%*	100%*
2. getting in/out of bath or shower	67.3%	76.9%	62.5%
3. sitting	83.3%	80.6%	70.8%
4. running	37.5%	42.1%	66.7%
5. twisting/pivoting on your loaded leg	87.5%*	85%*	54.2%
Comprehensiveness	31.7%	51.2%	16.7%

*: considered relevant, since ≥85% patients rated the item relevant. Pre-op: preoperative; Post-op: postoperative.

The item *degree of difficulty descending stairs* was indicated as relevant. Patients and experts commented that this item would not be relevant for all patients because not everyone uses stairs. However, most patients thought this item was relevant because they experience difficulty descending stairs due to their hip problems. Most of the experts commented that descending stairs is discriminative for the level of physical functioning and provides insight into the results of treatment (Figure 3.2b).

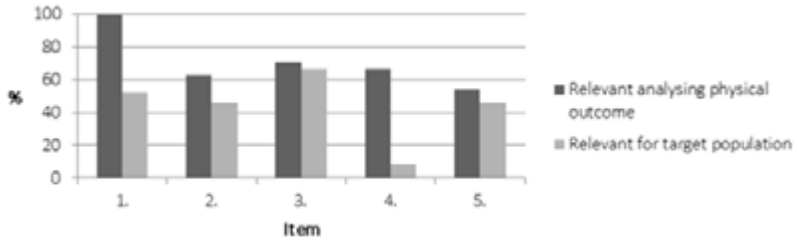

Figure 3.2a. Evaluation of the items of the HOOS-PS according to experts: relevancy of the items for the construct and population.
Items: Difficulty experienced with… 1. Descending stairs; 2. Getting in/out of bath; 3. Sitting; 4. Running; 5. Twisting/pivoting on loaded leg

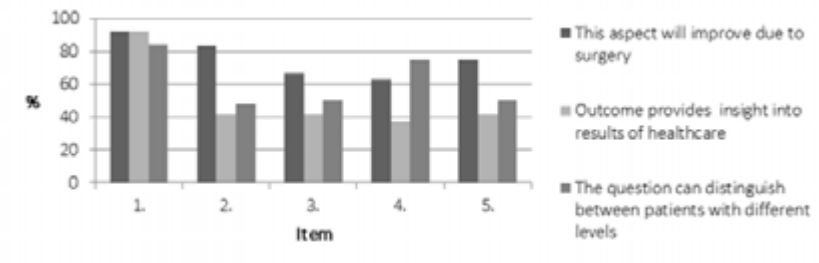

Figure 3.2b. Evaluation of the items of the HOOS-PS according to experts: responsiveness of scores on measurement and consequences in patient management.
Items: Difficulty experienced with… 1. Descending stairs; 2. Getting in/out of bath; 3. Sitting; 4. Running; 5. Twisting/pivoting on loaded leg.

The item *degree of difficulty getting in/out of bath or shower* was not considered relevant. Both respondent groups commented that this item would not be relevant for all patients because many patients do not have a bath. A patient explains: *"*I don't have a bath... Showering is simple, all I have to do is step into the shower... it is flat surfaced." Several patients who did use a bath, considered this item as relevant. A patient pointed out*:* "When I used the bath in a hotel, I could hardly get out due to my hip problems." Experts commented that a part of the question (difficulty getting in bath) is not relevant for the entire population. However, another expert noted: "it's an important activity of daily living."

The third item, *degree of difficulty sitting*, was not considered relevant by patients and experts. One expert commented: "As hip osteoarthritis will not often lead to a range of motion limitation of more than 90 degrees, sitting is not discriminative for the level of physical functioning." Experts thought the transfer (getting in and out of a seated

position) could be a more relevant measure of physical functioning. However, sitting was indicated as a common activity.

The item *degree of difficulty running* was considered not relevant for evaluating physical functioning by many patients and experts. The majority of the patients commented that the item was not applicable for their situation, because they do not run. This was confirmed by experts; only 8% of the experts thought that it would be relevant for patients undergoing THA. One expert states: *"*running is relevant for a different patient population, for instance for patients with hip labral tears." However, 75% of all experts found the item discriminative for good physical functioning.

The last item, *degree difficulty with twisting/pivoting on a loaded leg,* was found relevant by patients for evaluating physical functioning in patients with hip problems and not relevant by experts. Both patients and experts indicated that rotating on the operated leg may not be recommended after arthroplasty. Although the rotation of the hip is often affected by hip osteoarthritis, one expert stated that this problem gives no insight in the results of treatment because this activity rarely presents as a problem by patients when consulting the physician.

All of the patients reported that the response options of the measurement instrument and the recall period (1 week) were adequate. Therefore, the relevance including the items (insufficient relevance), response options and recall period (adequate relevance) was found inconsistent (Table 3.1).

Comprehensiveness

Comprehensiveness was found insufficient, meaning the HOOS-PS did not cover all aspects of physical functioning (Table 3.3). Patients identified concepts that were missing such as: walking, cycling, driving a car, sit to stand, standing and lying in bed. Less frequently, kneeling, squatting, putting on socks/shoes, playing sports and gardening were mentioned. Experts mentioned partly the same missing concepts (walking, cycling, playing sports, sit to stand transfer) and furthermore mentioned: bed transfer, climbing stairs, working, sexual activity and dressing.

Comprehensibility

The comprehensibility of the HOOS-PS was found insufficient. Although the instruction and words were understood by all patients and interpreted well, there were troubles with the comprehensibility of the items. The first item *descending stairs,* was well

interpreted by most patients. Only one patient interpreted the item as climbing stairs in general instead of specifically descending stairs. The majority of the patients suggested that the second item *getting in/out of bath or shower* was confusing. Talking about this issue, a patient said: "Stepping in/out of the bath is totally different than a walk-in shower" and filled in two options, none and severe difficulty. Experts judged the question also as double-barrelled, and suggested that getting in or out the shower requires much less physical capacity than getting in or out a bath. The third item *difficulty experienced during sitting* was interpreted as if the item measured how much pain was experienced. For example, one patient said: "If I sit on a dining chair, my hip is painful. However, it depends on the seat height." Experts commented: "What means having trouble sitting? Pain during sitting? Or physical inability?". Thus, a recurrent train of thought was that the item was questioning another construct, namely pain instead of physical functioning. The fourth item *difficulty running* was understood and well interpreted by all patients. The last item *twisting/pivoting on loaded leg* turned out hard to interpret. A response of a patient to this item was as follows: "They have to describe if the unaffected or the affected leg is meant." The experts agreed that this question was ambiguous. Other patients questioned which activity the item was linked to, for example*:* "twisting/pivoting on a loaded leg… for example dancing or something?" It can be concluded that the patient understanding of the item was doubtful. A suggestion to improve the clarity, was to link the item *twisting/pivoting on a loaded leg* to an activity.

The response options were indicated as appropriate by the majority of the patients and were consistent with the questions. However, some patients noted that a "not applicable" response option was missing. The recall period was regarded as appropriate by all patients. One patient concluded: "my symptoms vary from day to day; however I can adequately recall last week."

Interpretability

There were no floor or ceiling effects, since none of the patients had the highest possible score (worse physical outcome) and five patients (11%) scored the lowest possible score (best physical outcome) post-surgery. In 19.6% of the patients, item 4 *difficulty experienced running* was left blank. The percentage of missing data for the other items was lower (range 0-6%). The paper-based HOOS-PS was indicated as an adequate mode for collecting data by all interviewed patients, although one patient mentioned that he would prefer an online survey.

Discussion

The present study was the first study to assess the content validity of the HOOS-PS in order to identify if this measurement instrument adequately reflects physical functioning in patients with hip problems. This study showed that the content validity of the HOOS-PS in patients undergoing hip arthroplasty is problematic in three areas: 1. lack of relevance; 2. insufficient comprehensiveness; and 3. insufficient comprehensibility of the items. These findings challenge the applicability of the HOOS-PS in clinical practice, research, VBHC and benchmarking. The key aspects will be discussed under three headings: relevance, comprehensiveness and comprehensibility.

Relevance

This study showed problems regarding the relevance of the majority of the items of the HOOS-PS, supported by a previous study in which *sitting* and *getting in/out of bath/shower* were found not relevant[19]. Moreover, in this study the item *running* was also assessed not relevant. Several explanations were found, for example, physical incapability of doing the activity (e.g. running), the absence of the object mentioned in the item (e.g. stairs or bath), and the absence of experienced difficulty doing the activity (sitting, twisting/pivoting on loaded leg). The most likely cause of the inconsistent relevance is the lack of patient involvement in the development of the questionnaire[4]. Since four out of five items seem not to measure physical functioning of patients with hip problems (getting in/out bath or shower, sitting, running, twisting/pivoting on loaded leg), this finding suggest that the HOOS-PS is not an adequate measure of physical functioning.

The response options and recall period were found adequate. While patients did suggest a "not-applicable" response option, the addition of a non-applicable option is problematic. In case an item is left blank, the score of the HOOS-PS cannot be calculated since missing data are not accepted. Especially because some items were considered difficult to comprehend, more missing data are expected when adding a not-applicable response option. If the HOOS-PS continues to be used, agreements must be made to deal with these missing data. A possible solution is to migrate to a digital mode of administration, where questions can be made mandatory to complete. Last, the scoring rate can be adjusted, and items left blank get a predetermined score.

Comprehensiveness

When shortening a PROM, comprehensiveness is especially of interest. The HOOS-PS is a short-form and intends to represent the whole domain of physical functioning. The results of this study show that the HOOS-PS has an insufficient comprehensiveness, i.e. the majority of patients and experts thought the HOOS-PS did not cover all aspects of physical functioning. It is not possible to evaluate whether this is a consequence of the shortening, because the comprehensiveness of the full version is unknown. Respondents considered that there were a few missing key concepts. For example, respondents identified "walking" as a missing key concept. However, in the development article this item was explicitly excluded by Rasch analysis[4]. Furthermore, patients suggested adding several other items to the HOOS-PS, e.g. pain during the night, pain rotating the hip or stiffness. Although pain and stiffness are important for this population, they are not part of the intended measured construct (physical functioning) of the HOOS-PS and should not be included in the measurement instrument. Although comprehensiveness is important, balancing the number of items between redundancy and omissions is needed to obtain an optimum between burden and comprehensiveness.

Comprehensibility

This study showed that the instruction and response options were understood by patients as intended and the items were appropriately worded. Furthermore, the items *descending stairs* and *running* were considered comprehensible. However, the item *getting in/out of bath or shower* was found to be double-barrelled, the item *twisting/pivoting on loaded leg* hard to interpret and ambiguous and the item *sitting* was interpreted as questioning another construct (pain instead of physical functioning). Double-barrelled or ambiguous questions can be misinterpreted by patients and therefore can increase measurement error and can subsequently lead to incorrect conclusions.

Interpretability

There was a high amount of missing data of the item *running* in this study. This supports the finding that the item is not relevant, since missing data can suggest that the item is not relevant for the target population[14]. The absence of floor and ceiling effects does not imply that the items reflect the total range of ability of physical functioning. First, it seems that the items do not reflect the construct of physical functioning. Furthermore, in a large amount of patients the floor and ceiling effects were not analysed, since the scores could not be calculated in case of missing items.

Limitations

This study assessed the comprehensibility of the Dutch version of the HOOS-PS. Comprehensibility may be variable per language since there can be nuances in translations and not all translated versions were validated. Furthermore, the content validity in patients undergoing THA may be different in patients with other hip problems.

The cut-off value for considering an item relevant was arbitrary (>85%). A high cut-off achieves more certainty that the desired construct is measured and indicates that the items are relevant for the target population. To capture the full range of physical functioning it is possible that some items are relevant for only a small number of patients. However, since the HOOS-PS consists of only five items, it is debatable whether lowering the cut-off value is appropriate. Lowering the cut-off value for assessing an item as relevant, causes a reduction in the number of items relevant for one patient (with an already short measurement instrument).

A limitation of the development study with implications for this study, is that the development study did not base the construct of physical functioning on a conceptual or theoretical model. Although the developers aimed to measure physical functioning, it is noteworthy that the items of this questionnaire seem to measure only one domain of physical functioning, namely mobility. However, the definition physical functioning is conceptually multidimensional: "with four related subdomains: mobility (lower extremity function), dexterity (upper extremity function), axial (neck and back function), and ability to carry out instrumental activities of daily living."[13]. While not every subdomain is relevant for hip osteoarthritis (i.e. dexterity), it is important that the subdomains included in the theoretical model are mentioned by the developers so the content validity can be evaluated. Future studies developing PROMs should be based on an adequate theoretical framework, this is important for evaluating content validity in the desired context.

Strengths

The COSMIN standards were followed to achieve a high-quality study[14] in which open-ended questionnaires were complemented by in-depth interviews. A wide range and high number of patients undergoing total hip arthroplasty were included, indicating a heterogeneous sample, so the findings of this study may be able to be extrapolated to all total hip arthroplasty patients. The same accounted for the expert group which was multidisciplinary, including orthopaedic surgeons, physical therapists and orthopaedic surgery residents.

Conclusions

The use of the HOOS-PS is widespread in the evaluation of physical functioning in patients undergoing arthroplasty. This is the first report on the content validity of the HOOS-PS in a representative sample of patients undergoing THA. This study showed that the content validity of the HOOS-PS in patients undergoing hip arthroplasty is problematic in three areas: 1. lack of relevance; 2. insufficient comprehensiveness; and 3. insufficient comprehensibility of the items. These findings challenge the applicability of the HOOS-PS in clinical practice, research, VBHC and benchmarking.

Article and author information

Author contributions

Concept and design: Braaksma, Wolterbeek, Veen, Prinsen, Ostelo
Acquisition of data: Braaksma, Wolterbeek, Veen
Analysis and interpretation of data: Braaksma, Wolterbeek, Veen, Prinsen, Ostelo
Drafting of the manuscript: Braaksma, Wolterbeek, Veen
Critical revision of the paper for important intellectual content: Wolterbeek, Veen, Prinsen, Ostelo
Statistical analysis: Braaksma, Wolterbeek, Veen, Prinsen, Ostelo
Provision of study materials or patients: Braaksma, Wolterbeek, Veen
Administrative, technical, or logistic support: Braaksma, Wolterbeek, Prinsen, Ostelo
Supervision: Wolterbeek, Veen, Prinsen, Ostelo

Conflict of interest disclosures

The authors reported no conflicts of interest.

Funding/support

The authors received no financial support for this research.

Acknowledgment

The authors thank Dr C.B. Terwee, Department of Epidemiology and Data Science at Amsterdam University Medical Centre, for critically appraising and improving this manuscript.

References

1. Mokkink LB, Terwee CB, Patrick 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;63(7):737-745. 2. Rolfson O, Wissig S, van Maasakkers L, et al. Defining an International Standard Set of Outcome Measures for Patients With Hip or Knee Osteoarthritis: Consensus of the International Consortium for Health Outcomes Measurement Hip and Knee Osteoarthritis Working Group. *Arthritis Care Res*. 2016;68(11):1631-1639.
3. Peters RM, van Beers LWAH, van Steenbergen LN, et al. Similar Superior Patient-Reported Outcome Measures for Anterior and Posterolateral Approaches After Total Hip Arthroplasty. *J Arthroplasty*. 2018;33(6):1786-1793.
4. Davis AM, Perruccio A V., Canizares M, et al. The development of a short measure of physical function for hip OA HOOS-Physical Function Shortform (HOOS-PS): an OARSI/OMERACT initiative. *Osteoarthr Cartil*. 2008;16(5):551-559.
5. Braaksma C, Wolterbeek N, Veen MR, Prinsen CAC, Ostelo RWJG. Systematic review and meta-analysis of measurement properties of the Hip disability and Osteoarthritis Outcome Score - Physical Function Shortform (HOOS-PS) and the Knee Injury and Osteoarthritis Outcome Score - Physical Function Shortform (KOOS-PS). *Osteoarthr Cartil*. 2020;28(12):1525-1538.
6. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: A synthesis of recommendations. *Acad Med*. 2014;89(9):1245-1251.
7. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): A 32-item checklist for interviews and focus groups. *Int J Qual Heal Care*. 2007;19(6):349-357.
8. Gossec L, Hawker G, Davis AM, et al. OMERACT/OARSI initiative to define states of severity and indication for joint replacement in hip and knee osteoarthritis. *J Rheumatol*. 2007;34(6):1432-1435.
9. 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;15(12):1833-1840.
10. Klässbo 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;32(1):46-51.
11. de Groot IB, Reijman M, Terwee CB, et al. Validation of the Dutch version of the Hip disability and Osteoarthritis Outcome Score. *Osteoarthr Cartil*. 2007;15(1):104-109. doi:10.1016/j.joca.2006.06.014
12. World Health Organization. Towards a Common Language for Functioning , Disability and Health ICF. *Int Classif*. 2002;1149:1-22. http://www.who.int/classifications/icf/training/icfbeginnersguide.pdf.
13. Riley WT, Rothrock N, Bruce B, et al. Patient-reported outcomes measurement information system (PROMIS) domain names and definitions revisions: Further evaluation of content validity in IRT-derived item banks. *Qual Life Res*. 2010;19(9):1311-1321.
14. Terwee CB, Prinsen CAC, Chiarotto A, et al. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. *Qual Life Res*. 2018;27(5):1159-1170.
15. Terwee CB, Prinsen CA, Chiarotto A, et al. COSMIN methodology for assessing the content validity of PROMs: User manual. www.cosmin.nl. Published 2018. Accessed October 1, 2021.
16. Hak T, van der Veer K, Jansen H. The Three-Step Test-Interview (TSTI): An observation-based method for pretesting self-completion questionnaires. *Surv Res Methods*. 2008;2(3):143-150.
17. Schildmann EK, Groeneveld EI, Denzel J, et al. Discovering the hidden benefits of cognitive interviewing in two languages: The first phase of a validation study of the Integrated Palliative care Outcome Scale. *Palliat Med*. 2016;30(6):599-610.
18. van Wilgen CP, Verhagen EALM. A qualitative study on overuse injuries: The beliefs of athletes and coaches. *J Sci Med Sport*. 2012;15(2):116-121.
19. Nilsdotter AK, Lohmander LS, Klässbo M, et al. Hip disability and osteoarthritis outcome score (HOOS) - Validity and responsiveness in total hip replacement. *BMC Musculoskelet Disord*. 2003;4:1-8.

PART II

Towards an adequate alternative patient-reported outcome measure in THA and TKA

CHAPTER 4

Assessing the measurement properties of PROMIS Computer Adaptive Tests, short forms and legacy patient reported outcome measures in patients undergoing total hip arthroplasty

Journal of Patient-Reported Outcomes 2024 Oct 21;8(1):121

Abstract

Background

The commonly used ('legacy') PROMs evaluating outcomes of total hip arthroplasty (THA), have several limitations regarding their measurement properties and interpretation of scores. One innovation in PROMs is the use of Computerized Adaptive Testing (CAT). The Patient-Reported Outcomes Measurement Information System (PROMIS®) is a validated system of CATs. The aim of this study was to assess the measurement properties of PROMIS and legacy instruments in patients undergoing THA.

Methodology

Patients in this multicenter study filled out a questionnaire twice, including Dutch-Flemish PROMIS v1.2 Physical Function (PROMIS-PF) and v1.1 Pain Interference (PROMIS-PI) CATs and short forms, PROMIS v1.0 Pain Intensity, and legacy PROMs (Hip disability and Osteoarthritis Outcome Score (HOOS), HOOS-Physical function Shortform (HOOS-PS), Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), Oxford Hip Score (OHS), and two numeric rating scales measuring pain). The reliability, measurement precision (Standard Error of Measurement (SEM)), smallest detectable change (SDC), and burden of PROMIS instruments were presented head-to-head to legacy PROMs. Furthermore, construct validity was assessed.

Results

208 patients were included. All instruments had a sufficient test-retest reliability (range ICC: 0.83–0.96). The SEM of PROMIS CATs and short forms ranged from 1.8 to 2.2 T-score points, the SEM of legacy instruments 2.6–11.1. The SDC of PROMIS instruments ranged from 2.1 to 7.3 T-score points, the SDC of legacy instruments 7.2–30.9. The construct validity of PROMIS CAT and short forms were found sufficient, except for the PROMIS-PI short form. The burden of PROMIS CATs was smaller than PROMIS short forms (range 4.8–5.2 versus 8–20 items, respectively). The burden of legacy instruments measuring physical functioning ranged from 5 to 40 items.

Conclusions

The PROMIS-PF is less burdensome, with high measurement precision, and almost no minimal or maximal scores, and an equal reliability compared to legacy instruments measuring physical functioning in patients undergoing THA. The PROMIS Pain Intensity 1a is comparable to the legacy pain instruments in terms of burden, reliability and SDC. Measuring the construct Pain Interference may not have additional value in this population because of its high correlation with instruments measuring physical functioning. The SDC values presented in this study can be used for individual patient monitoring.

Introduction

Patient reported outcome measures (PROMs) are questionnaires, aiming to obtain information about perceived symptoms and functioning of the patient. PROMs are increasingly used in clinical practice to screen and monitor patient's symptoms and functioning, to facilitate informed and shared-decision making, and to improve quality-of-care[1]. However, much is still unknown about the optimal application of PROMs in daily clinical practice.

Total hip arthroplasty (THA) is number four ranked most frequently performed inpatient surgical procedure in the USA[2]. The use of PROMs in healthcare requires reliable and valid PROMs, with as little burden as possible for the patient. Unfortunately, the commonly used (called 'legacy') PROMs evaluating outcomes of THA, have several limitations regarding their measurement properties and interpretation of scores[3-6]. For example, the measurement error is often too large for reliable use of PROMs for individual patients, questions are often not relevant for all patients or not at all time points, and there is a lack of responsiveness, thereby hampering the ability to measure treatment effects[3-5]. Lastly, many of these PROMs have a limited measurement range causing floor and ceiling effects[3-5]. In conclusion, the legacy PROMs are not optimal for individual clinical assessment.

One promising innovation in PROMs is the use of Computerized Adaptive Testing (CAT). CAT can be used with PROMs that are developed using Item Response Theory (IRT) modelling[7]. IRT item banks are large sets of questions that are ordered in terms of their difficulty on an underlying metric. Using CAT, the most informative questions from item banks are selected depending on previous answers given by patients, until a predefined reliability is reached[8]. Patients are more likely to answer only relevant questions because e.g., questions about running will not be asked if a patient answers that he has difficulty walking one mile. Patients need to complete on average only four to seven questions to get a reliable score[9]. The use of CAT will decrease patient burden and, since the item banks cover the full width of the domain, floor and ceiling effects are less likely. The Patient-Reported Outcomes Measurement Information System (PROMIS®) is the most carefully developed and extensively validated system of CATs for measuring health outcomes[10], and is increasingly used in orthopedic clinical practice[11-16]. In addition to CAT, all PROMIS measures are also available as static short forms, containing a fixed subset of questions from the item bank. The short form scores are expressed on the same metric (scale) as scores obtained through CAT, and therefore, directly comparable.

These short forms could be administered when CAT is not (yet) technically possible within the data collection system of a clinic.

Using a PROM in individual clinical care is only helpful when the clinician and patient can interpret the score, and more specifically the change score over time. If the clinician or the patient is interested if a change in score is a real change (not due to measurement error), it is important that the Smallest Detectable Change (SDC) of the measurement instrument is known. The SDC is defined as the smallest change that can be detected by the instrument, beyond measurement error. There is little published data regarding the smallest detectable change (SDC) of PROMIS Physical Function or Pain Interference in the orthopedic field[17]. Furthermore, the theoretical benefit of PROMIS CAT and short forms administering patient friendly and relevant questionnaires, need to be confirmed in the clinical setting. Therefore, measurement properties of PROMIS CAT and short forms have to be determined presented head-to-head with the legacy PROMs in patients undergoing arthroplasty to investigate if PROMIS CAT and short forms overcome the limitations of the legacy PROMs.

The aim of this study was to assess and present the reliability, measurement precision, smallest detectable change, and burden of the Dutch-Flemish PROMIS Physical Function and Pain Interference CATs and short forms, and PROMIS Pain intensity head-to-head to legacy PROMs (the Hip disability and Osteoarthritis Outcome Score (HOOS), the HOOS-Physical function Shortform (HOOS-PS), Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC v3.1), Oxford Hip Score (OHS), and two numeric rating scales measuring pain at rest and pain during activity) in patients undergoing THA. Furthermore, construct validity of PROMIS CATs and short forms was assessed.

Methods

The study involved three orthopedic departments with high volumes of THAs in the Netherlands (St. Antonius Hospital Utrecht, Kliniek ViaSana Mill, OLVG Amsterdam). The study was conducted according to the principles of the Declaration of Helsinki. The study was reviewed by a Medical Ethics Review Committee (MEC-U) (St. Antonius Hospital, Nieuwegein, the Netherlands) (W21.037), which confirmed that the Medical Research Involving Human Subjects Act (WMO) does not apply. With this waiver, approval of the Institutional Review Board of each participating center was obtained.

Study participants

To ensure variability in PROM scores and to increase generalizability of the study results, two cohorts of patients were asked to participate: (1) patients currently on the waiting list for a THA and (2) patients who already underwent surgery. The patients in the second cohort were included at 3, 6 or 12 months post-surgery. As a rule of thumb, a sample size of 100 is considered as very good for the assessment of measurement properties[18]. To be eligible, patients had to be 18 years or older, and on the waiting list for a primary THA or 3, 6 or 12 months post-surgery. Exclusion criteria were THA for femoral neck fracture, patients unable to independently fill out questionnaires, insufficient knowledge of the Dutch language, or no internet facilities. Furthermore, patients who had surgery between test and retest were excluded. If patients were eligible and willing to participate, they were asked to sign the informed consent form digitally using an online informed consent module. Each hospital included a minimum of 25 patients, distributed over the measurement points.

Procedure

Patients were asked to fill out an online questionnaire twice within a two-week interval through a web-based platform (OnlinePROMS, Interactive Studios, 's-Hertogenbosch, the Netherlands). This is a certified (ISO27001; NEN7510), online PROMs platform, which is linked to the CAT software of the Dutch-Flemish Assessment Center, part of the Dutch-Flemish PROMIS National Center. A two-week interval was chosen to ensure no (large) changes in pain and function, which is a design requirement for assessing reliability, including smallest detectable change. A maximum of two automatic reminders were sent every two days after the first invitation when the patient had not responded. After two reminders the patient was considered lost-to-follow-up.

Measures

The questionnaire included two Dutch-Flemish PROMIS CATs, five Dutch-Flemish PROMIS short forms, one single PROMIS pain item, and six legacy PROMs. The retest questionnaire included the same questionnaires. The online platform did not allow for any missing values within questionnaires. Two PROMIS CAT measures were included: PROMIS v1.2 CAT Physical Function (PROMIS-PF) and PROMIS v1.1 CAT Pain Interference (PROMIS-PI; Table 4.1). The PROMIS CATs use a T-score metric with a mean of 50 and SD of 10, where 50 represents the mean score of the general population.

Table 4.1 Characteristics of included measurement instruments.

Questionnaire	Construct / definition	Items	Response options	Score	Recall	Reference
PROMIS measures						
PROMIS CAT Physical Function (PROMIS-PF, v1.2)	Functioning of one's upper extremities (dexterity), lower extremities (walking or mobility), and central regions (neck, back), as well as instrumental activities of daily living	Min 3 Max 12	5-point Likert	T-score [1]	-	[19,20]
PROMIS CAT Pain Interference (PROMIS-PI, v1.1)	Consequences of pain on relevant aspects of one's life	Min 3 Max 12	5-point Likert	T-score[1]	Last 7 days	[21]
PROMIS Physical Function SF8b (v1.2)	Functioning of one's upper extremities (dexterity), lower extremities (walking or mobility), and central regions (neck, back), as well as instrumental activities of daily living	8	5-point Likert	T-score[1]	-	[22,23]
PROMIS Physical Function SF10a (v1.2)	Functioning of one's upper extremities (dexterity), lower extremities (walking or mobility), and central regions (neck, back), as well as instrumental activities of daily living	10	5-point Likert	T-score[1]	-	[22,23]
PROMIS Physical Function SF20a (v1.2)	Functioning of one's upper extremities (dexterity), lower extremities (walking or mobility), and central regions (neck, back), as well as instrumental activities of daily living	20	5-point Likert	T-score[1]	-	[22,23]
PROMIS Pain Interference SF8a (v1.1)	Consequences of pain on relevant aspects of one's life	8	5-point Likert	T-score[1]	Last 7 days	[24,25].
PROMIS Pain Intensity 1a (v1.0)	How much a person hurts	1	11-option numeric rating scale	0 (no pain) - 10 (worst thinkable pain)	Last 7 days	[26,27].

Table 4.1 (continued).

Questionnaire	Construct / definition	Items	Response options	Score	Recall	Reference
Disease specific legacy PROMs						
Hip disability and Osteoarthritis Outcome Score (HOOS)	5 subscales - Pain - Symptoms - Stiffness - Function in daily living (ADL) - Function in sport and recreation (Sport/Rec) - Hip related Quality of Life (QOL)	40 - 10 - 3 - 2 - 17 - 4 - 4	5-point Likert	0 (indicating extreme symptoms) - 100 (indicating no symptoms)	Last week	[28]
HOOS- Physical Function Short form (HOOS-PS)	Physical functioning	5	5-point Likert	Raw scores were converted (0-100, 0 indicating extreme symptoms)[29]	Last week	[30]
Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC)	3 subscales: - Pain - Stiffness - Function	24 - 5 - 2 - 17	5-point Likert	Raw scores were converted (0-100, 0 indicating extreme symptoms[29]	Last 48 hours	[31]
Oxford Hip Score (OHS)	function and pain	12	5-point Likert	0-48 (0 indicating the worst, 48 the best outcome)	Past 4 weeks	[32]
NRS Pain activity	Pain during activity	1	11-option numeric rating scale	0-100 (0 indicating the worst, 100 the best outcome)	Last week	No reference available
NRS Pain rest	Pain at rest	1	11-option numeric rating scale	0-100 (0 indicating the worst, 100 the best outcome)	Last week	No reference available

[1] T-score 50 represents the average score of the general population, SD of 10

A higher PROMIS T-score represents more of the concept being measured (i.e. better function or more pain). The items in the CAT were selected based on their statistical ability to best further refine the individual's score, estimated from the already administered items. The CATs were automatically stopped when a Standard Error (SE) of 2.2 (95% reliability) was reached or a maximum of 12 items was administered. The CAT software used a Maximal Likelihood estimation (which was experimentally used for a while in the Netherlands with permission from HealthMeasures), in which in absence of variation in answer patterns, the calculation of the T-score and SE could were imputated (in this study the assigned scores were 0 or 100). Whenever a score could not be calculated using the ML estimation, the output of the score was 0 or 100 and registered as a minimum or maximum score. Table 4.2 shows the number and percentage of the patients with a minimum or maximum score.

Moreover, PROMIS short forms were administered: one measuring Pain Interference (SF8a) and three measuring Physical Function (SF8b, SF10a, and SF20a). These short forms contain a fixed set of items (Table 4.1). Scores are expressed on the same metric (scale) as scores obtained through CAT and, therefore, directly comparable. Furthermore, the PROMIS v1.0 Pain Intensity item 1a (also called Global07) was included. This item is also included in the PROMIS v1.2 Scale Global Health, validated as a brief measure of health related quality of life[26,27]. Moreover, the questionnaire included disease specific legacy PROMs: the Hip disability and Osteoarthritis Outcome Score (HOOS), the HOOS-Physical function Shortform (HOOS-PS), Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC v3.1), Oxford Hip Score (OHS), and two numeric rating scales measuring pain at rest and pain during activity (Table 4.1). The HOOS-PS and WOMAC were derived from the HOOS. Finally, the questionnaire included demographic and clinical characteristics (e.g. sex, age, joint, side, date of surgery).

Table 4.2 The ICC, the mean SEM, SDC and burden, the percentage patients with minimum and maximum scores and the range of PROMIS CAT, PROMIS short forms and legacy instruments (n=208).

	ICC agreement (CI)	SEM mean (range)	SDC mean (range)	Burden (mean) number of items	Minimum score (%)	Maximum score (%)	Score range
PROMIS-PF	.91(.88-93)	2.2(1.7-3.5)	6.9(4.7-9.6)	5.2	0%	0.7%	20.3-74.1
PROMIS- PI	.91(.87-.93)	2.1(1.9-5.9)	6.8(5.2-16.4)	4.8	13.2%	0.2%	44.6-76.8
PROMIS PF SF8b	.96(.95-.97)	2.2(1.5-5.9)	5.1(4.2-16.4)	8	0%	0%	20.9-59.7
PROMIS PF SF10a	.93(.91-.95)	2.2(1.7-5.9)	6.6(4.7-16.4)	10	0%	0%	20.9-61.9
PROMIS PF SF20a	.95(.94-.96)	1.8(1.3-5.7)	5.5(3.6-15.8)	20	0%	0%	20.6-62.7
PROMIS PI SF8a	.94(.92-.95)	2.4(1.3-5.9)	7.3(3.6-16.4)	8	0%	0%	40.7-77
PROMIS Pain Intensity 1a	.95(.93-.96)	0.8	2.1	1	18.3%	0.5%	0-10
	ICC agreement (CI)	**SEM**	**SDC**	**Burden number of items**	**Minimum score (%)**	**Maximum score (%)**	**Score range**
HOOS- PS	.83(.78-.88)	9.7	26.9	5	9.6%	0.5%	0-100
HOOS	.95(.93-.96)	6.3	17.6	40	0%	2.2%	1.9-100
HOOS- Symptoms	.91(.89-.93)	8.3	22.9	5	0.2%	9.8%	0-100
HOOS- QOL	.95(.93-.96)	7.5	20.9	4	9.4%	9.1%	0-100
HOOS-Sport/Recr	.88(.84-.91)	11.1	30.9	4	5.7%	7.4%	0-100
HOOS- ADL	.92(.89-.94)	7.5	20.7	17	0%	6.7%	1.5-100
HOOS- Pain	.93(.91-.95)	7.9	22	10	0.2%	15.3%	0-100
OHS	.96(.94-.97)	2.6	7.2	12	0%	7.2%	5-48
WOMAC	.92(.90-.94)	7.6	21	24	0%	5.4%	3.1-100
WOMAC - Pain	.90(.87-.92)	9.1	25.2	5	0.5%	22%	0-100
WOMAC - Stiffness	.87(.83-.90)	10.5	29.2	2	2.9%	13.2%	0-100
WOMAC - Function	.92(.89-94)	7.9	22	17	0%	6.7%	1.5-100
NRS pain Activity	.93(.91-95)	9.2	25.4	1	18.6%	2.2%	0-100
NRS pain Rest	.92(.90-.94)	8.5	23.6	1	29.4%	0.1%	0-100

Abbreviations: ICC= Intra-class Correlation Coefficient; SDC= smallest detectable change; SEM= Standard Error of Measurement; CI= confidence interval; QOL= quality of life; Sport/Recr = sports/recreation; ADL = activities of daily living; NRS = numeric rating scale; PI= Pain Interference; PF= Physical FunctionOutcomes

Reliability

Test-retest reliability

The test-retest reliability of the PROMIS CATs, PROMIS short forms and the legacy instruments was assessed by calculating the intra-class correlation coefficient (ICC) for each total- and/or subscale separately. Patients were invited twice within a two-week interval and, therefore, considered stable.

Measurement precision

The Standard Error of Measurement (SEM) at one time point was calculated as a parameter of measurement precision. PROMIS CATs and short forms were developed under an IRT model, in which each T-score is associated with its own standard error of measurement (SEM = SE(T-score)). The measurement error differs across the scale, each score (thus each patient) has its own SEM value. The legacy PROMs were developed under a Classical Test Theory (CTT) model, which assumes that all scores have the same SEM, so each PROM has one SEM value.

Smallest detectable change

Not every change on a measurement instrument can be considered a 'real' change. Small changes may be due to measurement error. The test-retest data were used to calculate the smallest detectable change (SDC), which is the smallest change in score that can be considered a 'real' change, above measurement error. The SDC is defined as the amount of change above which there is at least 95% chance that a real change has occurred[33].

Validity

Construct validity

Construct validity is defined as the degree to which the scores are consistent with hypotheses based on the assumption that the PROM validly measures the construct to be measured[34]. Hypotheses were formulated a priori about the expected correlations between the PROMIS CAT and PROMIS short forms with the comparator legacy instruments per measured domain. Correlations with measurement instruments measuring the same construct (e.g. PROMIS-PF and the HOOS-PS) were expected to be strong. Also, the PROMIS CAT and SF Pain Interference should highly correlate with comparator instruments measuring physical functioning, according to previous research in patients with musculoskeletal conditions and pain (e.g. in patients with chronic pain[20], spinal pain[35], and foot and ankle conditions[36]). This is expected because when pain levels increase, an individual's physical function decreases. Furthermore, the correlations with measurements instruments measuring the same construct, should be higher than measurement instruments measuring different but related constructs (e.g. PROMIS-PF and legacy measures of pain, stiffness or quality of life).

Interpretability

Burden

The number of items (also referred to as 'burden') needed to asses physical functioning and pain was compared between the PROMIS CAT, PROMIS short forms and the legacy PROMs.

Range of scores

Per measurement instrument, the percentage of patients with the minimal and maximum possible score were described.

Statistical analysis

Reliability

Test-retest reliability

The ICC was calculated using a two-way random-effects model for absolute agreement: $ICC\ agreement = \frac{\sigma_p^2}{\_^2 \cdot \_^2 \cdot \_^2}$, whereby σ_p^2 is the variation between patients, σ_m^2 is the variation between measurements and σ_e^2 is random error variance. Test-retest reliability was considered sufficient if ICC ≥0.70[33].

Measurement precision

The SEM for the legacy PROMs was calculated from the formula: $SEMagreement = \sqrt{\sigma^2{}_m + \sigma^2{}_e}$. Focusing on the absolute agreement, the variation between measurements (indicating systematic differences) is also considered error variance. The SEM (SE(T-score)) was provided for each patient score automatically when using PROMIS CAT software. For interpretation purposes, the mean and range of the SEM values were calculated and presented. There is no widely accepted method to compare the SEM or SDC of measurement instruments with different underlying theories (CTT versus IRT), since they have different scales. Therefore, the mean and the range presented can be used to interpret the corresponding measurement instrument and to compare measurement instruments on the same scale.

Smallest detectable change

The SDC is calculated as $SDC = 1.96 * \sqrt{2} * \mathbf{SEM}$. For PROMs that use IRT-based scoring, the individual SEM of the test T-score and the individual SEM of the re-test T-score were used ($SDC = 1.96 * \sqrt{SE_1^2 + SE_2^2}$). For traditional PROMs this will result in one SDC value (because there is only one SEM) per PROM, while for PROMs that use

IRT-based scoring, this will result in a different SDC for each patient. Therefore, the mean and the range (T-scores) are presented per measurement instrument.

Validity

Construct validity

To assess construct validity, Pearson's correlations were calculated between the PROMIS CAT and short forms, and the legacy PROMs. A matrix with all predefined hypotheses, resulting in 91 unique hypotheses, is presented in Supplemental Table S4.1. Construct validity was considered sufficient if ≥75% of the results was in accordance with the hypotheses.

Results

In total, 208 patients were included in the analyses (Figure 4.1). The mean age of the patients was 67.6 years, 62.8% were female (n=130). The mean time-interval between test and retest was 8 days (SD 2). The mean score, standard deviation and range per measurement instrument at different time points can be found in Supplemental Table S4.2.

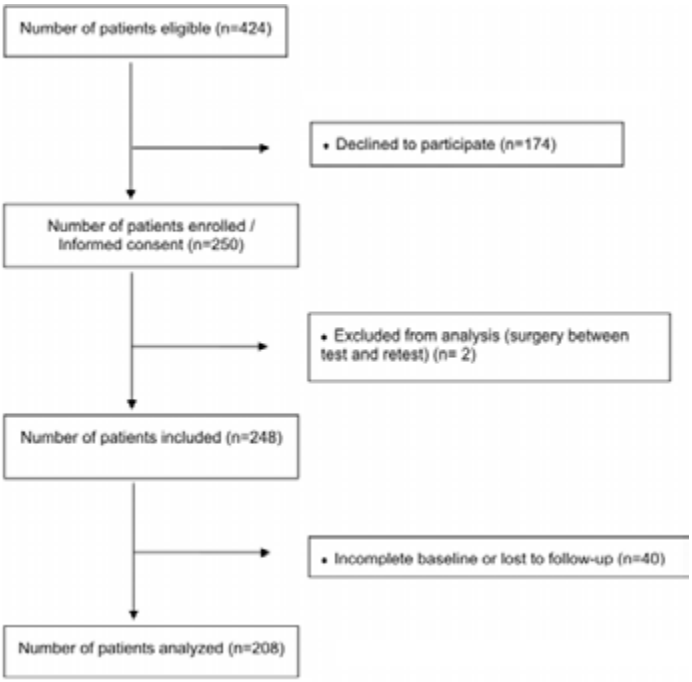

Figure 4.1. Flowchart of inclusion.

Reliability

Test-retest reliability

All PROMIS CATs, PROMIS short forms and legacy instruments showed evidence of sufficient test-retest reliability (range ICC: .83-.96, Table 4.2).

Measurement precision

The mean SEM of PROMIS CAT and short forms was 1.8-2.2 on the T-score scale (observed score range 20.3-77; Table 4.2). The SEM of PROMIS Pain intensity was 0.8 (score range 0-10). The SEM of the legacy instruments varied between 6.3-11.1 of legacy instruments with a score range 0-100, and was 2.6 for the OHS (observed score range 5-48; Table 4.2). The possible range of the instruments can be found in Table 4.1, the range of the observed scores are presented in Table 4.2. The distribution of the scores, the SEM and the SDC are presented in Figure 4.2.

Smallest detectable change

Table 4.2 gives details of the smallest detectable change of all PROMIS CATs, short forms and legacy instruments. The value of the SDC of PROMIS instruments was 2.1-7.3 T-score points (observed score range 20.3-77). The SDC of PROMIS Pain Intensity was 2.1 (score range 0-10). The SDC of the legacy instruments varied between 17.6-30.9 of legacy instruments with a score range 0-100. The SDC of the OHS was 7.2 (observed score range 5-48).

Validity

Construct validity

The construct validity of PROMIS CAT and short forms measuring Physical Function were sufficient (92.3 -100% of the results were in accordance with the hypotheses). The construct validity of the PROMIS Pain Intensity single item and PROMIS-PI were also found sufficient (both 92.3% of the results were in accordance with the hypotheses). The construct validity of PROMIS short form measuring Pain Interference was found insufficient (69.2% of the results were in accordance with the hypotheses) (Supplemental material Table S4.3).

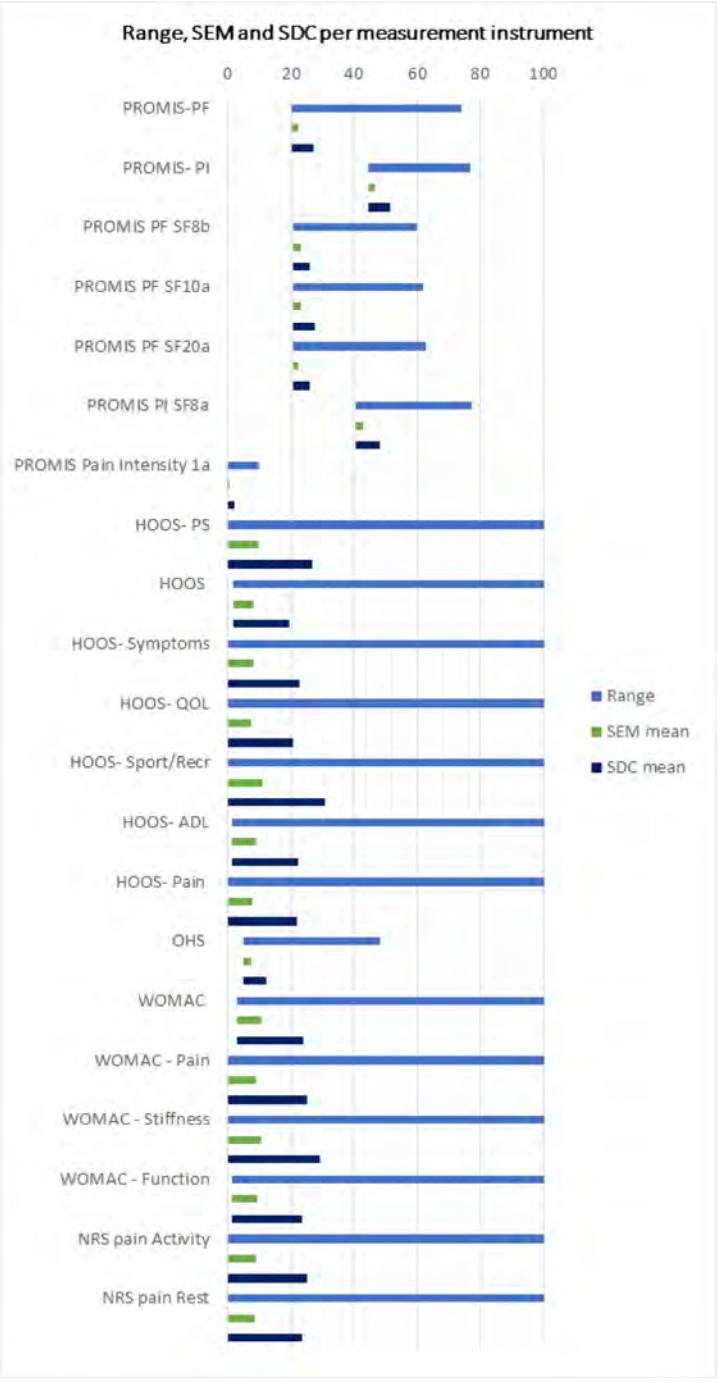

Figure 4.2. Range, SEM and SDC per measurement instrument.

Abbreviations: SEM= Standard Error of Measurement; SDC= smallest detectable change

Interpretability

Burden

The number of items administered per measurement instrument are presented in Table 4.2. The burden of PROMIS-PF and PROMIS-PI was smaller than PROMIS short forms (4.8-5.2 versus 8-20 items). The burden of legacy instruments measuring physical functioning varied between 5-40 items.

Range of scores

Table 4.2 shows the percentage of patients with the minimal and maximum score per measurement instrument. Witch exception of the PROMIS Pain Interference SF8a, all measurement instruments measuring pain had a considerable percentage of patients with a minimal or maximum score (13-18% of the scores of PROMIS instruments, 15-29% of the scores of legacy instruments). None of the patients had the minimum or maximum value on the PROMIS short forms measuring Physical Function. Less than 1% of the patients had a maximum score on the PROMIS-PF.

Discussion

The aim of this study was to determine if PROMIS CATs and short forms overcome the limitations of the legacy PROMs, by investigating the reliability, measurement precision, smallest detectable change, and burden of PROMIS Physical Function and Pain Interference CATs and short forms, and PROMIS Pain intensity, head-to-head to legacy PROMs in patients undergoing THA.

A clinically relevant finding is that PROMIS CATs are less burdensome with an equal reliability compared to legacy instruments in patients undergoing THA. Furthermore, this study reported on the SDC of many frequently used measurement instruments for patients undergoing THA. These SDC values per measurement instrument can be used as a guide to select a PROM with low measurement error, or as cut off values in the outpatient clinic to determine if it is likely that a patient has changed as result of the treatment.

This study faced methodological challenges in comparing the SEM and SDC between PROMIS and legacy instruments. The SEM and SDC can be used to interpret the measurement error of the measurement instruments and to compare measurement error of measurement instruments on the same scale. Although the absolute (mean) values of the SDC of PROMIS instruments were smaller than those of the legacy instruments, they cannot be directly compared, since measurement instruments have different scales (Figure 2.). The SDC is a value that represents the change that can be detected with 95% confidence on the scale of the corresponding measurement instrument. However, scales differ in unit of measurement (score on a specific legacy instrument or T-score), range and level of measurement (ordinal versus interval). Legacy instruments are developed using CTT (in which each item contributes equal to the score) and PROMIS measurement instruments using IRT (each item has its own difficulty and a weighted score is used). IRT implies that PROMIS instruments have equal intervals between values (i.e., interval scale) and legacy instruments don't (ordinal scale). To our knowledge, there is no consensus on how to address this problem. Several methods have been used in the literature to bypass this problem. One approach is to express the scores of different measurement instruments on the same IRT scale[37]. However, this method does not take into account that legacy instruments are not developed using IRT modelling. Other authors compared the percentage improved patients beyond measurement error, according to PROMIS and according to the legacy instruments[38]. Another solution would be to compare only ICC values, which relate the measurement error to the variation in scores. ICC values of the PROMIS measures were mostly higher than those of the legacy instruments. Because of the mentioned difficulties, this study presents the values per measurement instrument, accompanied with corresponding scales. More research is needed to determine the best approach to compare the measurement error of CTT-based and IRT-based instruments.

It should be noted that a lower CAT SE (SE 2.2, comparable to a reliability of 0.95) was used as stopping rule than the standard (SE 3.0, comparable to a reliability of 0.90). More reliable outcome scores can ensure more accurate individual patient monitoring, improve reliability of study results and can contribute to increase the use of patient reported outcome measures in the consultation room[1]. However, by using this setting it is presumable that the burden of the CATs increase (although in this study they were still lowest of all measurement instruments).

This study found that the PROMIS CAT and SF measuring Pain Interference were highly correlated with the comparator instruments measuring physical functioning in this patient population (resp. Pearson's r = .82; .87). These correlations were even higher

than the correlations with legacy instruments measuring pain. High correlations between PROMIS Physical Function and Pain Interference have also been found in previous studies[22,35,36], especially in patients suffering pain. It could be argued that for patients with pain these constructs are very similar. Because of this overlap in these constructs, it could be argued that there is no additional value measuring both in these patients. It could also be hypothesized that the construct pain is not relevant for all patient at every time moment, since most instruments measuring pain, had a considerable percentage of minimal or maximal possible scores, probably caused by the absence of pain post THA.

A possibly important aspect for THA patients when selecting the most suitable PROM for clinical practice is burden. Moreover, a smaller burden leads to less data storage, with subsequent reduction of the carbon footprint. A further reduction of the amount of data collected can be achieved by using PROMIS CAT and short forms, since these measurement instruments are generic and therefore the same PROMs can be used for multiple diagnoses.

When investigating alternatives for measuring physical functioning, the PROMIS-PF is less burdensome, has a wider measurement range (reducing floor/ceiling effects with more relevant questions) and almost no minimal or maximal possible scores, with an equal reliability compared to legacy instruments. When preferring a PROMIS Physical Function short form instead of PROMIS CAT, the 8-item PROMIS-PF SF 8b does not have a higher SEM or SDC than a short form containing more items (PF10a or PF20a). Furthermore, the 20-item PROMIS PF short form seems to add very little in score range beyond the PF 8b. Therefore, we recommend using the PROMIS-PF SF8b instead of PF10a or PF20A to reduce burden while obtaining an equal reliability and scoring range.

Regarding the construct pain, the PROMIS Pain Intensity 1a seems to be comparable to the legacy numeric rating scales measuring pain at rest and pain during activity in terms of burden, reliability and SDC. To facilitate the choice of an outcome measure, future research must focus on the minimally important change (MIC) and responsiveness of the different measures.

Conclusion

The PROMIS-PF is a less burdensome alternative, with a wider measurement range (reducing floor/ceiling effects with more relevant questions) and almost no minimal or maximal possible scores, with an equal reliability, compared to legacy instruments measuring physical functioning in patients undergoing THA. The PROMIS Pain Intensity

1a seems to be comparable to the legacy numeric rating scales measuring pain at rest and pain during activity in terms of burden, reliability, and SDC. Measuring the construct Pain Interference may not have additional value to measuring physical function in patients undergoing THA. The SDC and SEM of many frequently used measurement instruments presented in this study can be used as a guide to select a PROM, or as cut off values in the outpatient clinic to determine if it is likely that a patient has changed as result of the treatment.

Supplementary information

Acknowledgements

We want to thank Ariena Rasker for her contributions to the development of the design of the study and the data collection at the OLVG.

Author contributions

C.B. and N.W. conceptualized, arranged the acquisition, analyzed and did the management and coordination of the project. C.B., N.W., Y.P. and A.D.K. collected data. C.B., N.W., M.R.V., R.W.P., Y.P., A.D.K., R.W.J.G.O. and C.B.T. were involved in the design of the study. C.B. wrote the initial draft. N.W., M.R.V., R.W.P., Y.P., A.D.K., C.B.T. and R.W.J.G.O. reviewed, edited and supervised. All authors read and approved the final manuscript.

Funding

This research was funded by the LROI. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data availability

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

The study was conducted according to the principles of the Declaration of Helsinki. The study was reviewed by a Medical Ethics Review Committee (MEC-U) in the Netherlands, which confirmed that the Medical Research Involving Human Subjects Act (WMO) does not apply. With this waiver, approval of the Institutional Review Board of each participating center was obtained.

Consent for publication

Not applicable.

Competing interests

C.B.T. is a member of the PROMIS Health Organization and lead the Dutch- Flemish PROMIS National Center, which aim to improve health outcomes by developing, maintaining, improving, and encouraging the application of PROMIS in research and clinical practice. R.W.P. is a member of the Scientific and Innovation Committee (CWI) of the Dutch Orthopedic Association.

References

1. Amtmann D, Cook KF, Jensen MP, et al (2010) Development of a PROMIS item bank to measure pain interference. Pain 150:173–182. https://doi.org/10.1016/j.pain.2010.04.025
2. Baumhauer JF (2017) Patient-Reported Outcomes — Are They Living Up to Their Potential? N Engl J Med 377:6–9. https://doi.org/10.1056/nejmp1702978
3. Bellamy N, Buchanan WW, Goldsmith CH, et al (1988) 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 15:1833–1840
4. Bjorner JB, Chang CH, Thissen D, Reeve BB (2007) Developing tailored instruments: Item banking and computerized adaptive assessment. In: Quality of Life Research. pp 95–108
5. Braaksma C, Wolterbeek N, Veen MR, et al (2020) Systematic review and meta-analysis of measurement properties of the Hip disability and Osteoarthritis Outcome Score - Physical Function Shortform (HOOS-PS) and the Knee Injury and Osteoarthritis Outcome Score - Physical Function Shortform (KOOS-PS). Osteoarthr. Cartil. 28:1525–1538
6. Brodke DJ, Saltzman CL, Brodke DS (2016) PROMIS for Orthopaedic Outcomes Measurement. J Am Acad Orthop Surg 24:744–749. https://doi.org/10.5435/JAAOS-D-15-00404
7. Cella D, Riley W, Stone A, et al (2010) The patient-reported outcomes measurement information system (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. J Clin Epidemiol 63:1179–1194. https://doi.org/10.1016/j.jclinepi.2010.04.011
8. Collins NJ, Misra D, Felson DT, et al (2011) Measures of knee function: International Knee Documentation Committee (IKDC) Subjective Knee Evaluation Form, Knee Injury and Osteoarthritis Outcome Score (KOOS), Knee Injury and Osteoarthritis Outcome Score Physical Function Short Form (KOOS-PS), Knee Ou. Arthritis Care Res (Hoboken) 63 Suppl 1:S208-28. https://doi.org/10.1002/acr.20632
9. Copsey B, Thompson JY, Vadher K, et al (2019) Problems persist in reporting of methods and results for the WOMAC measure in hip and knee osteoarthritis trials. Qual. Life Res. 28:335–343
10. Crins MHP, Roorda LD, Smits N, et al (2015) Calibration and validation of the Dutch-Flemish PROMIS pain interference Item Bank in patients with chronic pain. PLoS One 10:. https://doi.org/10.1371/journal.pone.0134094
11. Crins MHP, Terwee CB, Klausch T, et al (2017) The Dutch–Flemish PROMIS Physical Function item bank exhibited strong psychometric properties in patients with chronic pain. J Clin Epidemiol 87:. https://doi.org/10.1016/j.jclinepi.2017.03.011
12. Crins MHP, van der Wees PJ, Klausch T, et al (2018) Psychometric properties of the PROMIS Physical Function item bank in patients receiving physical therapy. PLoS One 13:. https://doi.org/10.1371/journal.pone.0192187
13. Davis AM, Perruccio A V., Canizares M, et al (2008) The development of a short measure of physical function for hip OA HOOS-Physical Function Shortform (HOOS-PS): an OARSI/OMERACT initiative. Osteoarthr Cartil 16:551–559. https://doi.org/10.1016/j.joca.2007.12.016
14. Dawson J, Fitzpatrick R, Carr A, Murray D (1996) Questionnaire on the perceptions of patients about total hip replacement. J Bone Jt Surg - Ser B 78:185–190. https://doi.org/10.1302/0301-620x.78b2.0780185
15. De Ayala RJ (2009) The theory and practice of item response theory. Guilford Press
16. De Vet HCW, Terwee CB, Mokkink LB, Knol DL (2011) Measurement in medicine: A practical guide
17. Flens G, Terwee CB, Smits N, et al (2022) Construct Validity, Responsiveness, and Utility of Change Indicators of the Dutch-Flemish PROMIS Item Banks for Depression and Anxiety Administered as Computerized Adaptive Test (CAT): A Comparison With the Brief Symptom Inventory (BSI). Psychol Assess 2022 Jan; 34(1):58-69. https://doi.org/10.1037/pas0001068
18. Greenhalgh J, Gooding K, Gibbons E, et al (2018) How do patient reported outcome measures (PROMs) support clinician-patient communication and patient care? a realist synthesis. J. Patient-Reported Outcomes 15:2:42
19. Gupta P, Czerwonka N, Desai SS, et al (2023) The current utilization of the patient-reported outcome measurement information system (PROMIS) in isolated or combined total knee arthroplasty populations. Knee Surg. Relat. Res. 35(1):3

20. Hays RD, Bjorner JB, Revicki DA, et al (2009) Development of physical and mental health summary scores from the patient-reported outcomes measurement information system (PROMIS) global items. Qual Life Res 18(7):873-80:. https://doi.org/10.1007/s11136-009-9496-9
21. Hung M, Nickisch F, Beals TC, et al (2012) New paradigm for patient-reported outcomes assessment in foot & ankle research: Computerized adaptive testing. Foot Ankle Int 33:621–626. https://doi.org/10.3113/FAI.2012.0621
22. Kendall R, Wagner B, Brodke D, et al (2018) The relationship of PROMIS pain interference and physical function scales. Pain Med (United States) 19(9):1720-1724. https://doi.org/10.1093/pm/pnx310
23. Klässbo M, Larsson E, Mannevik E (2003) Hip disability and osteoarthritis outcome score: An extension of the Western Ontario and McMaster Universities Osteoarthritis Index. Scand J Rheumatol 32:46–51. https://doi.org/10.1080/03009740310000409
24. Lameijer CM, Van Bruggen SGJ, Haan EJA, et al (2020) Graded response model fit, measurement invariance and (comparative) precision of the Dutch-Flemish PROMIS® Upper Extremity V2.0 item bank in patients with upper extremity disorders. BMC Musculoskelet Disord 21(1):170. https://doi.org/10.1186/s12891-020-3178-8
25. Lawrie CM, Abu-Amer W, Barrack RL, Clohisy JC (2020) Is the Patient-Reported Outcome Measurement Information System Feasible in Bundled Payment for Care Improvement in Total Hip Arthroplasty Patients? J Arthroplasty 35:1179–1185. https://doi.org/10.1016/j.arth.2019.12.021
26. McDermott KW, Liang L (2021) Overview of Operating Room Procedures During Inpatient Stays in U.S. Hospitals, 2018. In: Healthc. Cost Util. Proj. Stat. Briefs [Internet]. Rockv. Agency Healthc. Res. Qual. (US); 2006 Feb-. Stat. Br. #281. Available from https//www.ncbi.nlm.nih.gov/books/NBK574416/
27. Mokkink LB, Terwee CB, Patrick DL, et al (2010) The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol 63:737–745. https://doi.org/10.1016/j.jclinepi.2010.02.006
28. Nilsdotter A, Bremander A (2011) Measures of hip function and symptoms: Harris Hip Score (HHS), Hip Disability and Osteoarthritis Outcome Score (HOOS), Oxford Hip Score (OHS), Lequesne Index of Severity for Osteoarthritis of the Hip (LISOH), and American Academy of Orthopedic Surgeons (AAOS) Hip and Knee Questionnaire. Arthritis Care Res 63 Suppl11:S200-7:. https://doi.org/10.1002/acr.20549
29. Nixon DC, McCormick JJ, Johnson JE, Klein SE (2017) PROMIS Pain Interference and Physical Function Scores Correlate With the Foot and Ankle Ability Measure (FAAM) in Patients With Hallux Valgus. Clin Orthop Relat Res 475(11):2775-2780. https://doi.org/10.1007/s11999-017-5476-5
30. Pellicciari L, Chiarotto A, Giusti E, et al (2021) Psychometric properties of the patient-reported outcomes measurement information system scale v1.2: global health (PROMIS-GH) in a Dutch general population. Health Qual Life Outcomes 19(1):226. https://doi.org/10.1186/s12955-021-01855-0
31. Prinsen CAC, Mokkink LB, Bouter LM, et al (2018) COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res. 27(5):1147-1157 https://doi.org/10.1007/s11136-018-1798-3
32. Quinzi DA, Childs S, Kuhns B, et al (2020) The Impact of Total Hip Arthroplasty Surgical Approach on Patient-Reported Outcomes Measurement Information System Computer Adaptive Tests of Physical Function and Pain Interference. J Arthroplasty 35:2899–2903. https://doi.org/10.1016/j.arth.2020.05.006
33. Rose M, Bjorner JB, Gandek B, et al (2014) The PROMIS Physical Function item bank was calibrated to a standardized metric and shown to improve measurement efficiency. J Clin Epidemiol.67(5):516-526 https://doi.org/10.1016/j.jclinepi.2013.10.024
34. Rothrock NE, Bass M, Blumenthal A, et al (2019) AO Patient outcomes center: Design, implementation, and evaluation of a software application for the collection of patient-reported outcome measures in orthopedic outpatient clinics. JMIR Form Res 3:(2):e10880. https://doi.org/10.2196/10880
35. Schuller W, Terwee CB, Klausch T, et al (2019) Validation of the Dutch-Flemish PROMIS Pain Interference Item Bank in Patients with Musculoskeletal Complaints. Spine (Phila Pa 1976) 44(6):411-419:. https://doi.org/10.1097/BRS.0000000000002847
36. Stephan A, Mainzer J, Kümmel D, Impellizzeri FM (2019) Measurement properties of PROMIS short forms for pain and function in orthopedic foot and ankle surgery patients. Qual Life Res 28:2821–2829. https://doi.org/10.1007/s11136-019-02221-w
37. Terwee CB (2020) The Value of Item Banks, CAT, and PROMIS for Dermatology. J Invest Dermatol 140:1089–1091. https://doi.org/10.1016/j.jid.2019.12.017

38. Terwee CB, Coopmans C, Peter WF, et al (2014) Development and validation of the computer-administered animated activity questionnaire to measure physical functioning of patients with hip or knee osteoarthritis. Phys Ther 94:251–261. https://doi.org/10.2522/ptj.20120472

Supplement

Table S4.1. Predefined hypotheses: expected correlations between the PROMIS CAT and PROMIS short forms Physical Function and Pain Interference, PROMIS single item Pain Intensity and legacy instruments.

	Physical functioning							Pain					Other domains		
PROMIS	OHS	HOOS-PS	HOOS ADL	HOOS sport	HOOS	WOMAC function	MEAN	NRS pain activity	NRS pain rest	WOMAC pain	HOOS pain	MEAN	HOOS symptoms	WOMAC stiffness	HOOS QOL
PROMIS-PF	A	A	A	A	A	A	$\bar{x}$	B	B	B	B		B	B	B
PROMIS-PI	A	A	A	A	A	A	$\bar{x}$	A	A	A	A	$\bar{x}$	B	B	B
PROMIS-PF SF8b	A	A	A	A	A	A	$\bar{x}$	B	B	B	B		B	B	B
PROMIS-PF SF10a	A	A	A	A	A	A	$\bar{x}$	B	B	B	B		B	B	B
PROMIS-PF SF20a	A	A	A	A	A	A	$\bar{x}$	B	B	B	B		B	B	B
PROMIS-PI SF8a	A	A	A	A	A	A	$\bar{x}$	A	A	A	A	$\bar{x}$	B	B	B
PROMIS Pain intensity 1a	B	B	B	B	B	B		A	A	A	A	$\bar{x}$	B	B	B

All correlations of PROMIS measurement instruments with the legacy measures were calculated. Measurement instruments measuring the same construct (e.g. PROMIS CAT Physical Function with WOMAC function subscale) were expected to have a Pearson's correlation >0.7 (hypotheses A). The mean of these correlations was calculated per PROMIS measure (column 'Mean', $\bar{x}$). Measurement instruments measuring related but different constructs were expected to have lower Pearson's correlation (hypotheses B) than the mean correlation of PROMs measuring the same construct ($\bar{x}$) (e.g. PROMIS CAT Physical Function correlation with NRS pain activity would be lower than the mean correlation of PROMIS CAT Physical Function with legacy instruments measuring physical function)

Table S4.2. Interpretability: average scores, range, presented as mean (SD; range).

	Pre-surgery (n=118)		**3 months post-surgery (n=38)**		**6 months post-surgery (n=24)**		**12 months post-surgery (n=28)**	
	Baseline	Retest	Baseline	Retest	Baseline	Retest	Baseline	Retest
PROMIS-PF	36.1(5.7; 20.3-51.6)	36.2(5.8; 22.2-55.2)	44.3(7.5; 25.2-61.6)	45.6(6.9; 25.2-56.2)	45(8.2; 27.8-57.6)	44.5(7.3; 27.5-57.6)	50.1(9.0; 32.6-74.1)	47.5(9.3; 21.9-63.9)
PROMIS-PI	65.5(5.5; 49.2-77.5)	65.2(5.9; 48.2-76.8)	56.2(8.7; 44.6-76)	55.6(9; 44.6-73.1)	54.4(8.7;44.6-73)	54.2(8.4; 44.6-75.4)	55(8.3; 44.6-69.2)	56(9.3; 44.6-74.2)
PROMIS-PF SF8b	34.6(5.8; 20.9-59.7)	34.6(5.7; 20.9-59.7)	43.1(8.6; 20.9-59.7)	44.9(8.6; 20.9-59.7)	44.7(8.8; 27.9-59.7)	44.1(8; 30.1-59.7)	50.7(9.1; 30.1-59.7)	49.7(9.1; 30.1-59.7)
PROMIS-PF SF10a	34.3(5.2; 20.9-53.4)	34(5.6; 20.9-51.2)	42.3(8.1; 26.9-61.9)	44.6(7.7; 32.5-61.9)	45.3(26-61.9)	44.7(8.1; 27.7-61.9)	49.8(9; 31.8-61.9)	49(9.9; 26-61.9)
PROMIS-PF SF20a	34.9(4.9;22.2-49.2)	34.2(5.5; 20.6-48.3)	43.1(7.6; 27.9-62.7)	44.9(7.6;29.2-62.7)	45.2(8; 30-62.7)	44.6(7.8; 31.2-62.7)	49.3(9; 32.4-62.7)	48.8(9.9; 25.8-62.7)
PROMIS-PI SF8a	63.6(5.6; 40.7-73.5)	63.8(6.2; 40.7-77)	51.7(9.3; 40.7-77)	50.2(9.2; 40.7-71)	48.5(10.3; 40.7-72.1)	49.7(9.4; 40.7-69.2)	47(8.6; 40.7-66.9)	47.5(9; 40.7-66.9)
PROMIS Pain intensity 1a	6.6(1.9; 0-10)	6.8(1.8; 0-10)	2.2(2.7; 0-9)	2(2.8; 0-8)	2.2(2.7; 0-9)	2.61(2.7; 0-8)	1.6(2.5; 0-8)	1.9(2.8; 0-9)
HOOS	33.2(16.4; 8.9-94.4)	37.1(17.2; 1.9-91.3)	77.2(19; 30-98.8)	79(18.6; 26.9-99.4)	81.8(17.9; 38.8-98.8)	76(22.8; 24.4-98.1)	89(14.3; 47.5-100)	88.2(14.5; 55-100)
HOOS-PS	46.1(17; 0-82.4)	51(19.1; 4.6-100)	20.2(15.4; 0-61.6)	20.4(18;0-82)	15.1(16.3; 0-67.9)	18.8(17.1; 0-61.6)	8.2(10.3; 0-30.4)	10.4(13.1; 0-46.1)
HOOS-Symptoms	40(19.4; 5-95)	38.6(17.7; 0-90)	76.6(18; 35-100)	78.1(16.9;35-100)	80(21.2; 35-100)	75.9(23.1; 25-100)	87.5(16.7; 50-100)	85(18.5; 35-100)
HOOS-QOL	23.9(18.8; 0-93.8)	24.2(19.1; 0-93.8)	67.4(24.8; 6.3-100)	69.3(24.5; 6.25-100)	77.2(23.7; 25-100)	73.6(25.6; 12.5-100)	82.4(24.9; 0-100)	82.8(23.5; 18.8-100)
HOOS-Sport/Recr	26.6(19.4; 0-87.5)	23(19.7; 0-87.5)	58.6(28.4; 0-100)	62.9(26.6; 0-100)	73.9(22.2; 25-100)	66.3(27.6; 0-100)	82.6(20; 43.8-100)	79.9(24.9; 12.5-100)
HOOS-ADL	44.8(17.7; 8.8-98.5)	41.6(18.8; 1.5-97.1)	80.5(18.9; 29.4-100)	81.5(19; 23.5-100)	83.8(16.5; 50-100)	77.2(25; 25-100)	91.1(12.7; 47.1-100)	91.1(12; 64.7-100)
HOOS-Pain	40.2(18.1; 5-100)	39.6(17.7; 0-90)	83.5(19; 35-100)	85.4(17; 40-100)	84.2(21.7; 27.5-100)	78.8(24.2; 25-100)	91.3(13.8; 47.5-100)	90.4(13.7; 52.5-100)
OHS	23.4(8.8; 5-47)	22.6(9.6; 3-47)	38.8(7.9; 14-48)	39.7(8; 14-48)	40.4(8.6; 10-48)	39.6(9.9; 9-48)	43.8(6.3; 19-48)	43.1(8.3; 8-48)
WOMAC	44.4(17.5; 9.4-99)	42.3(18.5; 3.1-96.9)	80.8(18.8; 31.3-100)	82.1(18.2; 26-100)	83.5(17.8; 40.6-99)	77.5(24.5; 26-99)	90.8(12.8; 47.9-100)	90.2(12.5; 64.6-100)
WOMAC-Pain	45.8(20.4; 0-100)	45.2(18.8; 0-100)	87(18.8; 35-100)	87.9(16.2; 35-100)	85.2(23.8; 25-100)	80.7(24.8; 25-100)	92.9(12; 55-100)	91.6(11.3; 65-100)
WOMAC-Stiffness	37.2(20.5; 0-100)	37.4(19.2; 0-87.5)	68.4(26.8; 0-100)	68.8(24.7; 0-100)	76.6(27.5; 0-100)	72.3(26.4; 12.5-100)	83.5(22.1; 25-100)	79.5(26.2; 0-100)
WOMAC-Function	44.8(17.7; 8.8-98.5)	41.6(18.8; 1.5-97)	80.5(18.9; 29.4-100)	81.5(19; 23.5-100)	83.8(16.5; 50-100)	77.2(25; 25-100)	91.1(12.7; 47-100)	91.1(12; 64.7-100)
NRS Pain Activity	73.4(18.1; 0-100)	72.2(19.2; 0-100)	24.5(27.9; 0-80)	19.4(23.8; 0-80)	24.5(27.9; 0-80)	24.4(28.1; 0-80)	12.14(19.1; 0-70)	11.8(21.6; 0-90)
NRS Pain Rest	53.7(23.3; 0-100)	54.3(22.9; 0-100)	11.6(16.7; 0-50)	11.9(18.5; 0-70)	15.2(22.9; 0-90)	15.2(26.6; 0-90)	8.6(19.4; 0-70)	8.6(18.6; 0-70)

Abbreviations: SD= standard deviation; n= number of patients; CI= confidence interval; SDC= smallest detectable change; QOL= quality of life; Sport/Recr= sports/recreation; ADL= activities of daily living, NRS= numeric rating scale

Table S4.3. Pearson's r for correlations between PROMIS CATs and short forms and legacy instruments (n=208).

	PF						Pain					Other				
Measurement instrument	OHS	HOOS-PS	HOOS ADL	WOMAC function	WOMAC total	MEAN	NRS pain activity	NRS pain rest	WOMAC pain	HOOS pain	MEAN	HOOS sport/rec	HOOS symptoms	WOMAC stiffness	HOOS QOL	% according to hypotheses
PROMIS-PF	.83	-.74	-.80	.80	.80	**.79**	-.63	-.72	.76	.79		.77	.77	.74	.78	**92.3**
PROMIS-PI	-.85	.76	-.82	-.82	-.83	**.82**	.70	.78	-.80	-.83	**.78**	-.75	-.77	-.73	-.82	**92.3**
PROMIS-PF SF8b	.84	-.79	.83	.83	.84	**.83**	-.67	-.76	.78	.81		.83	.80	.79	.82	**92.3**
PROMIS-PF SF10a	.84	-.81	.84	.84	.83	**.83**	-.70	-.75	.77	.80		.82	.80	.77	.81	**100**
PROMIS-PF SF20a	.85	-.81	.84	.84	.84	**.84**	-.70	-.76	.78	.80		.82	.81	.78	.81	**100**
PROMIS-PI SF 8a	-.89	.83	-.88	-.88	-.89	**.87**	.74	.84	-.86	-.89	**.83**	-.86	-.86	-.83	-.89	**69.2**
PROMIS Pain intensity 1a	-.86	.81	-.86	-.86	-.88		.86	.90	-.86	-.89	**.88**	-.80	-.84	-.82	-.86	**92.3**

Green= hypotheses accepted; red= hypotheses rejected. PF= Physical Function.

CHAPTER 5

A comparison of the psychometric properties of PROMIS Computer Adaptive Tests and short forms versus legacy patient-reported outcome measures in total knee arthroplasty patients

Arthroplasty Today 2026 38; in press.

Abstract

Background

Traditionally used patient-reported outcome measures (PROMs) evaluating total knee arthroplasty (TKA) have large measurement error and limited measurement range. PROMIS© has theoretically possibilities to overcome these limitations. This study evaluates the psychometric properties of PROMIS versus legacy PROMs in TKA patients.

Materials and methods

210 patients were included from three orthopaedic departments. Patients completed a questionnaire twice in a two-week interval, including two PROMIS Computer Adaptive Tests (CATs), assessing Physical Function (PF) and Pain Interference (PI), four PROMIS short forms (SF) (PF-SF8b, PF-SF10a, PF-SF-20a, PI-SF8a, PI-1a) and six legacy PROMs (KOOS, KOOS-PS, WOMAC, OKS, two numeric rating scales). Reliability, measurement precision, smallest detectable change (SDC), construct validity, burden and extreme scores were investigated.

Results

All PROMIS CATs, SFs and legacy PROMs showed adequate test-retest reliability. PROMIS CATs and SFs showed sufficient construct validity. Regarding physical function, the SDC varied between 2.8-6.1 for PROMIS, and 7.3-19.9 for legacy PROMs. PROMIS CAT and SFs showed no extreme scores using 5-20 items, legacy PROMs showed 1.2-3% extreme scores using 7-17 items. Concerning pain, the SDC varied between 2.8-6.1 for PROMIS and 22.5-35.4 for legacy PROMs. PROMIS showed 0-9.5% extreme scores using 5-8 items, legacy PROMs 8.7-19.1% using 1-9 items.

Conclusions

PROMIS CAT and SFs seem more efficient assessing physical function in TKA compared to legacy PROMs by offering reduced burden and measurement error, and minimizing extreme scores. PROMIS PI-8a seems most suitable in measuring pain without extreme scores. This study supports a shift from legacy PROMs toward PROMIS in TKA patients.

Introduction

Patients' perspectives on total knee arthroplasty (TKA) effectiveness are evaluated using patient-reported outcome measures (PROMs). The traditionally used ('legacy') PROMs evaluating TKA have several disadvantages regarding measurement properties. The measurement error is often too large for reliable use in individual patients[1,2]. Other challenges include limited reliability and responsiveness[2]. Therefore, these PROMS are not optimal for use in the consultation room with the individual patient. More reliable measurements can ensure better patient monitoring and improve quality of care[3].

An innovation in PROMs is the possibility of using Computerized Adaptive Tests (CAT). CAT selects items from item banks, which are collections of questions calibrated on an underlying measurement scale, using Item Response Theory (IRT) modelling. The Patient-Reported Outcomes Measurement Information System (PROMIS®) is the largest and most extensively validated system of CATs for measuring self-reported health[4]. PROMIS CAT is a dynamic form of testing that adjusts the items questioned to a patient based on their previous responses. This makes the process more responsive to the patient's specific health condition and minimizes the response burden[5,6]. Patients need to complete, on average, only 4-7 items to get a reliable score[7]. Furthermore, the item banks cover the entire construct width and therefore, floor and ceiling effects likely do not arise.

In addition to CAT, all PROMIS measures are also available in static short forms (SFs), containing a fixed set of items from the item bank. SFs may be used when CAT is not yet feasible. SF and CAT scores are directly comparable since the scores are expressed on the same metric (scale). These SFs outperformed legacy PROMs in patients with osteoarthritis regarding reliability and measurement range. However, PROMIS CAT performed superiorly over SFs for most measurement properties and burden[8,9].

PROMIS is extensively used in the orthopaedic field. Most studies showed a sufficient correlation between PROMIS and legacy PROMs[10]. However, little is published on the comparative psychometric properties between PROMIS and legacy PROMs, with only a few studies in TKA patients[10–12].

This study aimed to compare the performance of PROMIS CATs and PROMIS SFs to legacy PROMs, focusing on function and pain in TKA patients. The reliability, measurement precision, smallest detectable change, construct validity, extreme scores

and burden of legacy PROMs will be compared head-to-head to PROMIS CATs and PROMIS SFs in TKA patients.

Materials and methods

Design

This is a multicenter test-retest study at three high-volume TKA orthopaedic departments in the Netherlands (St. Antonius Hospital in Utrecht, Kliniek ViaSana in Mill, OLVG in Amsterdam). Institutional Review Board approval was obtained for each participating center.

Study participants

Adult patients were eligible for study inclusion either preoperatively, while on the waiting list for a primary TKA, or postoperatively. This ensures variability in scores. From the study start date, consecutive patients awaiting surgery attending their six- or twelve-month postoperative follow-up were included in the study. Therefore, each patient completed a single test-retest assessment. Excluded were patients unable to independently fill out questionnaires or without internet facilities. This study was nested in the regular PROM administration. Eligible patients were approached during routine PROM administration in each hospital and asked to digitally sign informed consent. A sample size of 100 patients was considered very good for assessing measurement properties[13]. Each hospital included a minimum of 50 patients distributed over the measurement points.

Procedure

Patients were asked to complete an online questionnaire twice at a two-week interval. This design ensures no (large) changes in pain and function, and enables the assessment of reliability, including the smallest detectable change. Questionnaires were emailed through a web-based platform (OnlinePROMS, Interactive Studios, ‘s-Hertogenbosch, the Netherlands). This certified (ISO27001; NEN7510) OnlinePROMs platform is linked to the Dutch-Flemish Assessment Center CAT software. Missing data were not allowed, as the system is designed to prevent incomplete submissions. A maximum of two automatic reminders were sent every two days after the first invitation when the patient did not respond.

Measurement instruments

Patients' physical function, pain intensity, and pain interference were measured using two Dutch-Flemish PROMIS CATs, five Dutch-Flemish PROMIS short forms, and six legacy PROMs (Table 5.1). The re-test questionnaire included the same PROMIS CATs, PROMIS SFs, and legacy PROMs.

Table 5.1. Characteristics of included measurement instruments.

	Construct / definition	Items	Response options	Score	Recall
PROMIS measures					
PROMIS CAT Physical Function (PROMIS PF, v1.2[13,14])	Functioning of one's upper extremities (dexterity), lower extremities (walking or mobility), and central regions (neck, back), as well as instrumental activities of daily living	Min 3 Max 12	5-point Likert	T-score[1], higher T-score indicating better PF	-
PROMIS Physical Function SF20a (v1.2[16,17])	Functioning of one's upper extremities (dexterity), lower extremities (walking or mobility), and central regions (neck, back), as well as instrumental activities of daily living	20	5-point Likert	T-score[1], higher T-score indicating better PF	-
PROMIS Physical Function SF10a (v1.2[16,17])	Functioning of one's upper extremities (dexterity), lower extremities (walking or mobility), and central regions (neck, back), as well as instrumental activities of daily living	10	5-point Likert	T-score[1], higher T-score indicating better PF	-
PROMIS Physical Function SF8b (v1.2[16,17])	Functioning of one's upper extremities (dexterity), lower extremities (walking or mobility), and central regions (neck, back), as well as instrumental activities of daily living	8	5-point Likert	T-score[1], higher T-score indicating better PF	-
PROMIS CAT Pain Interference (PROMIS-PI, v1.1[30])	Consequences of pain on relevant aspects of one's life	Min 3 Max 12	5-point Likert	T-score[1], higher T-score indicating more pain	Last 7 days
PROMIS Pain Interference SF8a (v1.1[15,18])	Consequences of pain on relevant aspects of one's life	8	5-point Likert	T-score[1], higher T-score indicating more pain	Last 7 days
PROMIS Pain Intensity 1a (v1.0[19,20])	How much a person hurts	1	11-option numeric rating scale	0 (no pain) - 10 (worst thinkable pain)	Last 7 days
Legacy PROMs					
Knee injury and Osteoarthritis Outcome Score (KOOS[21])	5 subscales - Pain - Symptoms - Function in daily living (ADL) - Sport and recreation Function (Sport/Rec) Quality of Life (QOL)	42 - 9 - 7 - 17 - 5 - 4	5-point Likert	0 (indicating extreme symptoms) - 100 (indicating no symptoms)	Last week

Table 5.1 (continued)

	Construct / definition	Items	Response options	Score	Recall
Legacy PROMs					
KOOS Physical function Shortform (KOOS-PS[22])	Physical function	7	5-point Likert	Raw scores were converted (0-100, 0 indicating extreme knee symptoms) (31)	Last week
Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC[23])	3 subscales: - Pain - Stiffness Function	24 - 5 - 2 - 17	5-point Likert	Raw scores were converted (0-100, 0 indicating extreme knee symptoms) (31)	Last 48 hours
Oxford Knee Score (OKS[24])	Function and pain	12	5-point Likert	0-48 (0 indicating the worst, 48 the best outcome)	Past 4 weeks
NRS Pain activity (No reference available)	Pain during activity	1	11-option numeric rating scale	0-100 (0 indicating the worst, 100 the best outcome)	Last week
NRS Pain rest (No reference available)	Pain at rest	1	11-option numeric rating scale	0-100 (0 indicating the worst, 100 the best outcome)	Last week

[1] T-score 50 represents the average score of the general population, SD of 10

The two PROMIS CAT measures included were the PROMIS v1.2 CAT Physical Function[14,15] and PROMIS v1.1 CAT Pain Interference[16]**.** PROMIS CAT identifies the most relevant questions from its item banks based on responses to both initial and follow-up questions. This process continues until a specified level of reliability is reached. The CATs automatically stopped when a standard error (SE) of 2.2 was achieved, corresponding to 95% reliability, or when a maximum of 12 items were answered. The CAT software used in this study utilized maximum likelihood estimation, which means that determining the T-score and SE requires variation in item responses. When patient responses were uniform on the CAT (all positive or all negative responses), the SE could not be estimated, leading to imputed T-scores (these scores were set to 0 or 100).

Furthermore, three PROMIS SFs measuring physical function were included (PROMIS Physical Function SF8b, SF10a and SF 20a[17,18]. In addition, one PROMIS SF measuring Pain Interference (SF8a[16,19]) and the single item PROMIS Pain Intensity 1a[20,21] were included.

PROMIS uses a T-score metric with an average score of 50 and a standard deviation of 10. A score of 50 reflects the average performance of the general population. A higher PROMIS T-score indicates a greater degree of the measured concept (such as better function or more pain).

The questionnaire included the following legacy PROMs for physical function and pain: the Knee injury and Osteoarthritis Outcome Score (KOOS[22]), KOOS-Physical function Shortform (KOOS-PS[23]), Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC[24]), Oxford Knee Score (OKS(25)) and two numeric rating scales (NRS) measuring pain during activity and pain in rest (Table 5.1). The legacy PROMs were developed under a Classical Test Theory (CTT) model. Characteristics of these legacy PROMs can be found in Table 5.1.

Finally, patient characteristics (sex, age and date of surgery) were collected.

Data analysis

Reliability, measurement precision (Standard Error of Measurement, SEM), smallest detectable change (SDC), construct validity, burden and extreme scores were compared between PROMIS CATs, PROMIS SFs and legacy PROMS.

Test-retest reliability

The intra-class correlation coefficient (ICC) was calculated to assess test-retest reliability for each of the PROMIS CATs, PROMIS SFs and legacy PROMs with total- or subscales. The ICC was calculated using a two-way random-effects model for absolute agreement, following the formula: $ICC\ agreement = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_m^2 + \sigma_e^2}$. In this formula σ_p^2 is the variation between patients, σ_m^2 is the variation between measurements and σ_e^2 is the random error variance. An ICC value of 0.70 or higher was considered indicative of 'sufficient' test-retest reliability[26].

Measurement precision

Measurement precision was assessed by calculating the Standard Error of Measurement (SEM). Since the possible range of the score differs per outcome measure and is important for interpreting the SEM, both the SEM and the range of the score are presented. The PROMIS CATs and SFs were created using an IRT model. In this framework, each T-score is associated with a standard error of measurement (SEM= SE(T-score)). Measurement error is not constant across the scale, implying that each score (and each consequently each patient) has a unique SEM value. The PROMIS CAT

software automatically computes the SEM for each patient's score. The legacy PROMs are developed under a CTT model, assuming all scores have the same SEM. The SEM for the legacy PROMs was calculated using the formula: $SEMagreement = \sqrt{\sigma^2{}_m + \sigma^2{}_e}$.

Smallest detectable change

Minor fluctuations in scores may reflect true changes in the patient's condition, or may result from measurement errors. The smallest detectable change (SDC) was calculated, defined as the minimum change in score above which there is at least a 95% chance that a real change has occurred[26]. The SDC was calculated using the formula: $SDC = 1.96 * \sqrt{2} * \mathrm{SEM}$. For PROMIS CATs and SFs (based on IRT) the individual SEM for both the test T-score and the re-test T-score was used in the calculation: $(SDC = 1.96 * \sqrt{SE_1^2 + SE_2^2})$. They yield a distinct SDC value for each individual patient. Consequently, the mean and range of T-scores are presented for each PROMIS CAT or SF.

Given the variations in measurement scales, there is no widely accepted approach for comparing SEM or SDC of measurement instruments based on differing underlying theories (IRT versus CTT). We described these difficulties in a previous study by our research group[28]. Because of these difficulties, this study reports the number of points of the SEM and SDC within the instrument's specific scale. The reported SEM and SDC cannot be directly compared between measurement instruments but are considered useful for interpreting scores of a given measurement instrument.

Construct validity

Construct validity is defined as the degree to which a measure's scores are consistent with hypotheses based on the assumption that the intended construct is validly measured by the PROM[27]. To examine construct validity of PROMIS CAT and PROMIS SFs, hypotheses were formulated a priori, based on expected correlations between the PROMIS CAT and PROMIS SFs, and legacy PROMs for each construct (physical function and pain). 1). A high correlation was expected between instruments measuring the same construct (e.g., PROMIS PF and KOOS-PS). 2). Moreover, a high correlation was expected between PROMIS CATs and SFs in the construct pain with legacy PROMs evaluating physical functioning. And vice versa; a high correlation was expected between PROMIS CATs and SFs in the construct physical function with legacy PROMs evaluating pain. It has been shown that as pain levels increase, an individual's physical functioning typically declines. This reinforces the anticipated strong correlation[15,28–30]. Additionally, it was expected that correlations among measurement instruments evaluating the same

construct (e.g. PROMIS PF and KOOS-PS) would exceed those among instruments assessing different but related constructs (e.g. PROMIS PF and legacy PROMs measuring pain, stiffness or quality of life).

To evaluate construct validity, Pearson's correlations were calculated between the PROMIS CATs and SFs, as well as legacy PROMs, resulting in a total of 91 unique predetermined hypotheses. Construct validity was considered adequate if at least 75% of the findings confirmed the proposed hypotheses[13]. If a correlation was greater than 0.7, the hypotheses was accepted.

Feasibility

Burden

Burden was defined as the number of items per instrument to asses physical functioning and pain.

Interpretability

Range of scores

Extreme scores were defined as minimum or maximum possible scores. The percentage of patients with extreme scores was reported for each measurement instrument.

Results

A total of 210 participants were included. Figure 5.1 presents the flowchart of inclusion. The mean age was 68.3 years (SD 7.7), ranging from 45 to 85 years, with 51.9% being men (n=109). The mean time interval between test and re-test was 8 days (SD 2). Table 5.2 provides details on the mean (SD, range) scores of all PROMIS CATs, PROMIS SFs, and legacy PROMs.

Test-retest reliability

Results showed evidence for sufficient test-retest reliability for all PROMIS CATs, PROMIS SFs and legacy PROMs (ICCs between .74 and .94, Table 5.3).

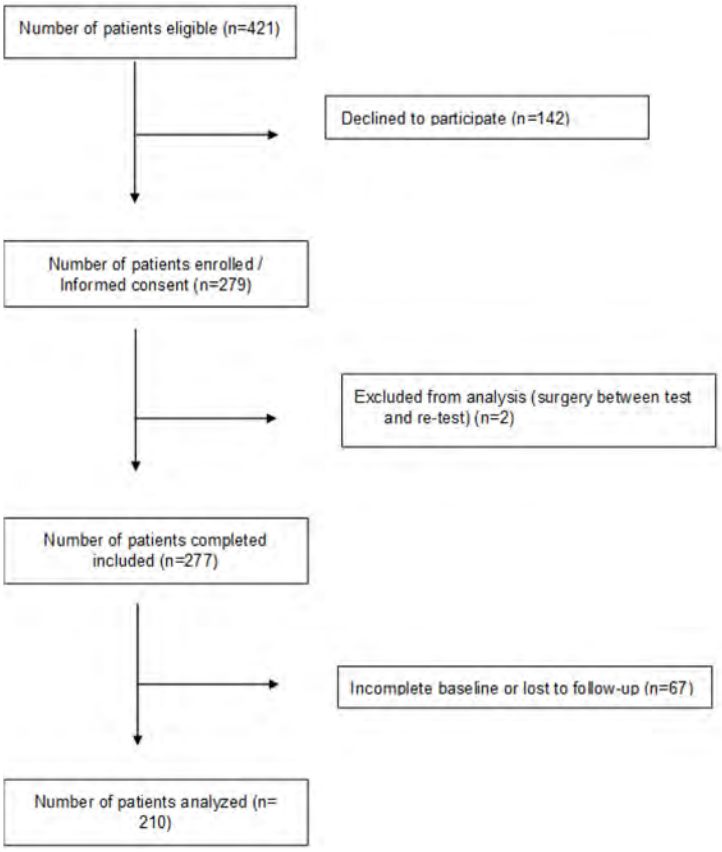

Figure 5.1. Flowchart of inclusion.

Measurement precision

The SEM of PROMIS CATs and SFs ranged between 1 and 2.1. The SEM of the legacy PROMs varied between 2.6 and 13.3. For legacy PROMs which are scaled from 0 to 100, the lowest SEM for pain and physical function was observed for respectively the KOOS pain (SEM 8.1) and the KOOS-PS (SEM 6.9) (Table 5.3).

Table 5.2. Average scores and range per measurement point.

	Mean (SD; range)					
	Pre-operative (n=115)		**6 months (n=67)**		**12 months (n=28)**	
	Baseline	**Re-test**	**Baseline**	**Re-test**	**Baseline**	**Re-test**
PROMIS CAT PF	37.4 (5.9; 18.7-53.1)	37.4 (6.4; 12.6-52.5)	45 (6.3; 23.8-54.9)	45.6 (6.3; 28.2-54.8)	45.1 (5.6; 34.4-53.6)	44.5 (5.9; 30.2-55)
PROMIS SF PF 20a	36.3 (5.5; 20.6-49.2)	36.2 (5.8;15.3-47.5)	43.1 (6.5;18.3-62.7)	42.9 (6.1;20.1-54.9)	43.3 (6.8;32.7-62.7)	43.9 (7.24;30.7-57)
PROMIS SF PF 10a	35.6 (5.3; 22-49.4)	35.7 (5.7;13.5 -47.9)	43.3 (6.7;20.9-61.9)	43 (6.2; 23.1-55.8)	43.1 (7.05;29.4-61.9)	43.9 (7.4;31-61.9)
PROMIS SF PF 8b	36.1 (5.6; 20.9-59.7)	36.1 (5.7;20.9-52.5)	43.8 (7.3; 20.9-59.7)	44.1 (7.4;24.4-59.7)	45.8(8.4;31.1-59.7)	45.1 (8.1;27.9-59.7)
PROMIS CAT Pain interference	63.6 (5.5; 50.4-81)	63.6 (5.4; 52-76.6)	55.4 (6.9;44.6-75)	54.7 (6;44.6-69.4)	55.7 (8.2;44.6-67.6)	55.9 (7;44.7-67.6)
PROMIS SF Pain interference 8a	62.1 (5.9; 40.7-77)	61.8 (5.7;48-77)	51.5 (8.1; 40.7-77)	50.2 (8.8; 40.7-68.4)	51.6 (8.9; 40.7-66.9)	49.9 (8.7; 40.7-64.1)
PROMIS SF Pain intensity 1a	6(2.1;1-10)	6.2 (2.1; 1-10)	2.7 (2.3;0-9)	2.5 (2.3;0-8)	2.3 (2.3; 0-7)	2.4 (2.2; 0-7)
KOOS	45.7 (17.3;4.2-82.1)	44.3 (16.9;2.4-82.7)	72.3 (16.5;5.4-100)	71.8 (15.7;31-100)	74.5 (16.2;36.9-94.6)	74.4 (16.6;44.6-100)
KOOS-PS	48.1 (16.1;14.8-100)	50.6 (16.7; 18.6-100)	30.1 (12.4; 0-62)	31.3 (12.1; 0-57.9)	29.5 (11.9; 10.5-51.2)	30.7 (14.2; 0-62)
KOOS-Symptoms	54.2 (21.7; 0-89)	52.8 (19.3; 0-96.4)	70.3 (17.1; 10.7-100)	71.1 (15.1; 39.3-100)	74.7 (17.2; 39.3-100)	76.4 (15.2; 46.4-100)
KOOS-QOL	29.6 (16.2; 0-81.3)	30 (15.9; 0-81.3)	61 (19.9; 0-100)	60.7 (19.3; 6.3-100)	64.3 (22.8; 6.3-100)	63.8 (22.6; 12.5-100)
KOOS-Sport/Recr	19.5 (17.1;0-65)	20.2 (24.1; 0-100)	44.9 (24.3; 0-100)	42.5 (27.4; 0-100)	45.5 (22.1; 5-85)	43.9 (27.9;0-100)
KOOS-ADL	53.2 (20.5; 5.9-91.1)	49.9 (19.9; 0-88.2)	80.6 (16.4; 7.4-100)	79.9 (15; 32.3-100)	82.2 (16.5;38.2-100)	81.3 (16.4; 48.5-100)
KOOS-Pain	46.9(19.7;0-83.3)	46.6 (18.8; 0-86.1)	78.6 (19.8; 0-100)	78.3 (17.8;19.4-100)	80.5 (18.3; 38.9-100)	81.7 (16.9; 47.2-100)
OKS	24.9 (8.1; 5-43)	24.3 (8.6; 1-42)	36.4 (8.3; 10-48)	36.6 (8.1; 12-47)	38.6 (7.9; 20-48)	38.4 (7.5; 24-48)
WOMAC	52.3 (19.9; 7.3-88.5)	49.4(19.1;0-87.5)	79.7(16.5; 5.2-100)	79(15.3; 32.3-100)	81.7 (16.5; 39.6-100)	81.2 (15.8; 49-100)
WOMAC-Pain	52.9 (21.2; 0-95)	50.9 (19.2; 0-90)	82.7 (18.5; 0-100)	81.6 (17.2; 30-100)	84.3 (18; 35-100)	85 (15.6;50-100)
WOMAC-Stiffness	43.7 (25.3; 0-100)	41.7 (22.5; 0-100)	64.4 (22.3;0-100)	61.8 (22.6; 12.5-100)	70.5 (22.4; 12.5-100)	71 (19.6; 25-100)
WOMAC-Function	53.2 (20.5; 5.9-91.2)	49.9 (19.9; 0-88)	80.6 (16.4;7.4-100)	79.9 (15; 32.4-100)	82.2 (16.5; 38.2-100)	81.3 (16.4; 48.5-100)
NRS Pain Activity	67.3 (22.4; 0-100)	65.9 (22.6; 0-100)	31.9(24.8; 0-80)	31(26; 0-90)	29.3(27.9;0-80)	26.1(25.3;0-80)
NRS Pain Rest	45.1 (25.2; 0-100)	45.4 (25.5;0-100)	20.5 (22; 0-70)	21.8 (24.4;0-90)	19.2 (24.2;0-80)	20 (21.6;0-70)

Abbreviations: SD= standard deviation; n= number of patients; MIC= minimally important change; CI= confidence interval; SDC= smallest detectable change; QOL= quality of life; Sport/Recr = sports/recreation; ADL = activities of daily living; SF= short form.

Smallest detectable change

The SDC for PROMIS CATs and SFs varied between 2.8 and 6.1 and for legacy PROMs between 7.3 and 37. Regarding physical function, the lowest SDC for PROMIS CAT and SF was found using the PROMIS PF SF20a (SDC 4.4). The lowest SDC among the legacy PROMs with a 0-100 scale was the KOOS-PS (SDC 19). Regarding the construct pain, the SDC of PROMIS CAT and SF8a was equal (SDC 6.1). The lowest SDC among the legacy PROMs with a 0-100 scale was the KOOS-pain (SDC 22.5). The SDC of the three numeric rating scales measuring pain (PROMIS Pain Intensity 1a and two legacy PROMs) was respectively 2.8 (scale 0-10), 33.9 and 35.4 (scale 0-100).

Table 5.3. The mean SEM, SDC, burden, ICC, and the percentage of patients with minimum and maximum scores of PROMIS CAT, PROMIS SFs and legacy PROMs.

	SEM mean (range)	SDC mean (range)	ICC agreement (95% CI)	Burden (mean) number of items	Minimum score (%)	Maximum score (%)	Score range
PROMIS instruments							
PROMIS CAT PF	2.1(1.9-2.2)	5.7(5.3-6.1)	0.90(0.87-0.92)	5.2	0	0	12.6-55
PROMIS PF SF20a	1.6(1.3-5.7)	4.4(3.6-13.2)	0.91(0.88-0.93)	20	0	0	15.3-62.7
PROMIS PF SF10a	2(1.7-5.9)	5.6(4.7-13.9)	0.89(0.86-0.92)	10	0	0	13.5-61.9
PROMIS PF SF8b	1.9(1.5-5.9)	5.4(4.2-16.4)	0.90 (0.87-0.92)	8	0	0	20.9-59.7
PROMIS CAT PI	2.1(1.7-3.6)	6.1(3.6-16.4)	0.78 (0.72-0.83)	4.8	7.7	0	44.6-81
PROMIS PI SF8a	2.1(1.3-5.9)	6.1(3.6-16.4)	0.84(0.80-0.88)	8	0	0	40.7-77
PROMIS Pain Intensity 1a	1	2.8	0.87(0.84-0.90)	1	9.3	0.2	0-10
Legacy PROMs							
KOOS	5.4	15.1	0.94(0.92-0.95)	42	0.3	0	2.4-100
KOOS- PS	6.9	19	0.85(0.80-0.88)	7	0.9	0.9	0-100
KOOS- Symptoms	7.2	19.9	0.88(0.85-0.91)	7	0.3	2.8	0-100
KOOS- QOL	7.1	19.7	0.91(0.89-0.93)	4	0.3	0	0-100
KOOS- Sport/Recr	13.3	37	0.74(0.67-0.80)	5	38.4	4.3	0-100
KOOS- ADL	7.2	19.9	0.91(0.88-0.93)	17	0.5	3.8	0-100
KOOS- Pain	8.1	22.5	0.89(0.86-0.92)	9	0.3	8.4	0-100
OKS	2.6	7.3	0.94(0.92-0.95	12	0	1.2	7.9-48
WOMAC	6.7	18.5	0.92(0.89-0.94)	24	0.2	2.1	0-100
WOMAC - Pain	8.4	23.4	0.88(0.85-0.91)	5	0.9	10.2	0-100
WOMAC - Stiffness	11.9	32.9	0.79(0.73-0.84)	2	2.8	6.4	0-100
WOMAC - Function	7.2	19.9	0.91(0.88-0.93)	17	0.2	2.8	0-100
NRS pain – Activity	12.2	33.9	0.83(0.79-0.87)	1	9.4	2.1	0-100
NRS pain - Rest	12.8	35.4	0.78(0.72-0.83)	1	18.4	.7	0-100

Abbreviations: SEM= Standard Error of Measurement; SDC= smallest detectable change; ICC= Intra-class Correlation Coefficient; CI= confidence interval; PF= Physical Function; PI= Pain Interference; QOL= quality of life; Sport/Recr = sports/recreation; ADL = activities of daily living; NRS = numeric rating scale; SF = short form.

Construct validity

The results indicated sufficient construct validity for all PROMIS CATs and PROMIS SFs (Table 5.4). 77 to 100% of the results were in accordance with the predefined

hypotheses. All measurement instruments measuring physical function (PROMIS CAT-PF, PROMIS SF PF 20a, 10a, and 8b) correlated highly (mean Pearson's r .81-.83) with legacy instruments measuring physical function and legacy instruments measuring pain (mean Pearson's r 0.71-0.73), as hypothesized. All instruments evaluating pain (PROMIS CAT- PI, PROMIS SF PI 8a, PROMIS Pain intensity 1a) correlated highly with legacy instruments measuring pain (mean Pearson's r 0.72-0.84) and physical function (mean Pearson's r 0.83-0.84). The correlations among the measurement instruments evaluating the same constructs exceeded those among instruments assessing different but related constructs.

Table 5.4. Pearson's r for correlations between construct of PROMIS CATs, PROMIS SFs and legacy PROMs (n=210).

Domain	PF						Pain					Other				TOTAL
Measurement instrument	OKS	KOOS-PS	KOOS ADL	WOMAC function	WOMAC total	*MEAN*[1]	NRS pain activity	NRS pain rest	WOMAC pain	KOOS pain	*MEAN*[1]	KOOS sport/rec	KOOS symptoms	WOMAC stiffness	KOOS QOL	% of findings aligned with hypotheses proposed hypotheses
PROMIS CAT Physical Functioning	**.86**	**-.74**	**.82**	**-.82**	**.82**	*.81*	**-.72**	-.59	**.77**	**.77**	*.71*	**.60**	**.62**	**.61**	.76	**85**
PROMIS SF Physical Functioning 20a	**.86**	**-.77**	**.84**	**.84**	**.83**	*.83*	**-.70**	-.59	**.76**	**.77**	*.71*	**.62**	**.66**	**.62**	.74	**85**
PROMIS SF Physical Functioning 10a	**-.86**	**.75**	**-.84**	**-.84**	**-.85**	*.83*	-.61	.65	**-.81**	**-.84**	*.73*	**-.64**	**-.67**	**-.65**	-.83	**77**
PROMIS SF Physical Functioning 8b	**.86**	**-.74**	**.83**	**.83**	**.83**	*.82*	**.73**	-.59	**.78**	**.78**	*.72*	**.61**	**.60**	**.62**	.78	**85**
PROMIS CAT Pain Interference	**-.86**	**.72**	**-.80**	**-.80**	**-.81**	*.84*	**-.71**	-.62	**-.76**	**-.78**	*.72*	**-.57**	**-.63**	**-.60**	-.82	**85**
PROMIS SF Pain interference 8a	**-.86**	**.75**	**-.84**	**-.84**	**-.85**	*.83*	**.77**	.65	**-.81**	**-.84**	*.77*	**-.64**	**-.67**	**-.65**	-.83	**85**
PROMIS SF Pain intensity 1a	**-.84**	**.75**	**-.84**	**-.84**	**-.85**	*.84*	**.87**	**.79**	**-.84**	**-.85**	*.84*	**-.61**	**-.64**	**-.65**	**.80**	**100**

Bold: hypotheses in line with expectation

[1:] The mean correlation was calculated per construct (physical function and pain) per measurement instrument. It was expected that the correlation per PROMIS instrument with the mean correlation of legacy PROMs measuring the same construct (such as PROMIS PF compared to the mean correlation of the legacy PROMs measuring physical function) would exceed those among instruments assessing different but related constructs (such as PROMIS PF compared to the mean correlation of legacy PROMs measuring pain, stiffness or quality of life).

Abbreviations: QOL= quality of life; Sport/Recr = sports/recreation; ADL = activities of daily living; SF= short form.

Feasibility

For the PROMIS CAT PF and PI, approximately five items were necessary to achieve the predefined reliability threshold (Table 5.3). Based on previous research, the average completion rate for PROMIS CAT instruments is approximately five items per minute[4]. The three PROMIS SFs for assessing Physical Function consisted of 8, 10 and 20 items respectively. The number of items of the legacy PROMs measuring physical function varied from 1–42. Regarding the construct pain, the included PROMIS SF measuring Pain Interference contained 8 items. PROMIS Pain Intensity contained a single-item. There were two single-item pain legacy PROMs included and two subscales of legacy PROMs measuring pain consisting of 5 and 9 items respectively.

Interpretability

PROMIS Physical Function CAT and SFs did not present any minimum or maximum scores. Legacy PROMs measuring physical function had 1.2-3.2% extreme scores. Regarding pain, 8.2% of the patients scored at the best end of the scale for the PROMIS PI CAT, 0% for the PROMIS PI SF8a and 9.5% for the PROMIS Pain Intensity. For legacy PROMs measuring pain, extreme scores were observed in 8.7 to 42.7% of patients (Table 5.2).

Discussion

This study showed that PROMIS CAT PF offers a more efficient alternative to legacy PROMs in assessing physical function in TKA patients, without compromising measurement quality. Both PROMIS CAT and SF measuring physical function avoid extreme score distributions. Regarding the construct pain, the PROMIS PI 8a was the only measurement instrument without extreme scores with an average burden of 8 items and sufficient reliability and construct validity.

PROMIS CATs and SFs, particularly CAT PF, may be better suited for clinical monitoring and decision-making in TKA patients due to their reduced burden lack of extreme scores, compared to legacy PROMs. The PROMIS PF CAT and SFs showed no extreme scores. Especially when there is a need to measure patients with extremely poor function or almost no symptoms, it is important that the instruments can measure at the extremes of the scale. These findings are consistent with those of Dhollander et al.[12], confirming that PROMIS PF demonstrates excellent precision, responsiveness, and feasibility in TKA patients[12]. Our multicenter design, test-retest reliability and inclusion of PROMIS Short

Forms and the Oxford Knee Score provide additional evidence on the broader applicability and comparative performance of PROMIS instruments. This study showed the superior performance of PROMIS CAT PF relative to PROMIS SF PF, previously described by Fries et al.[8]. However, when the implementation of CAT remains impractical the PROMIS PF SF8b is the best alternative. Although it consists of an average of three more items than the CAT, the SEM and SDC are comparable. Interestingly, the PROMIS SF20a has a slightly lower SDC and SEM. This is probably attributed to the higher number of items questioned, which is also the disadvantage of this SF.

As for the construct pain, only the PROMIS PI 8a could measure without extreme scores. All single item PROMs measuring pain (PROMIS Pain intensity 1a and both legacy PROMs) and the legacy PROMs were not good at measuring at the end of the scale (>8% extreme scores). Regarding the construct pain, PROMIS CAT and SF had a comparable SDC and SEM.

This study confirmed previously reported strong correlations between PROMs assessing pain and physical function[15,28–30]: higher pain levels correlate with lower levels of physical function. It could be hypothesized that measuring both pain interference and physical function might not be necessary in patients who experience pain.

The strength of this study was the thorough designed prospective multicenter study across the country, which may lead to generalizability of the results. A methodological limitation of our study is the use of the outdated Maximum Likelihood (ML) estimation, which may have contributed to a higher percentage of extreme scores in the PROMIS Pain Interference CAT. It is expected that when the standard EAP method will be used, there will be less extreme scores. Another methodological challenge was the interpretation of the SDC and SEM. This remains complex, as comparisons across different scales of CTT and IRT-based measurement instruments cannot be made adequately. We present the SDC and SEM values per measurement instrument, accompanied by the range of the scales for interpretation purposes. Other methods for bypassing this problem were previously described, but no consensus exist regarding the best approach[28].

To facilitate the selection of the most appropriate measurement instruments for clinical practice, future research should focus on comparing the minimal important change and responsiveness between PROMIS instruments and legacy PROMs in TKA patients. In

addition to pain and physical function, other domains as mental health could be assessed.

Conclusions

Both PROMIS CAT and SFs seem most efficient for assessing patient-reported physical function in TKA, compared to legacy PROMs. This by offering reduced burden and measurement error, and minimizing the occurrence of extreme scores. The PROMIS PI 8a seems most efficient for assessing pain in TKA patients, minimizing extreme scores. The results of this study can facilitate better patient monitoring and decision-making in TKA patients. These findings may suggest a need for a reevaluation of routine outcome measurement using legacy PROMs.

Acknowledgement

We want to thank Amanda Klaassen for her contributions to the development of the design of the study and the data collection at the OLVG.

References

1. Braaksma C, Wolterbeek N, Veen MR, Prinsen CAC, Ostelo RWJG. Systematic review and meta-analysis of measurement properties of the Hip disability and Osteoarthritis Outcome Score - Physical Function Shortform (HOOS-PS) and the Knee Injury and Osteoarthritis Outcome Score - Physical Function Shortform (KOOS-PS). Vol. 28, Osteoarthritis and Cartilage. 2020. p. 1525–38.
2. Gagnier JJ, Mullins M, Huang H, Marinac-Dabic D, Ghambaryan A, Eloff B, et al. A Systematic Review of Measurement Properties of Patient-Reported Outcome Measures Used in Patients Undergoing Total Knee Arthroplasty. Journal of Arthroplasty. 2017 32(5):1688-1697
3. Baumhauer JF. Patient-Reported Outcomes — Are They Living Up to Their Potential? N Engl J Med. 2017;377(1):6–9.
4. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. The patient-reported outcomes measurement information system (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. J Clin Epidemiol. 2010;63(11):1179–94.
5. Bjorner JB, Chang CH, Thissen D, Reeve BB. Developing tailored instruments: Item banking and computerized adaptive assessment. In: Quality of Life Research. 2007. p. 95–108.
6. Nguyen TH, Han HR, Kim MT, Chan KS. An introduction to item response theory for patient-reported outcome measurement. Vol. 7, Patient. 2014.
7. Segawa E, Schalet B, Cella D. A comparison of computer adaptive tests (CATs) and short forms in terms of accuracy and number of items administrated using PROMIS profile. Qual Life Res. 2020;29(1).
8. Fries JF, Cella D, Rose M, Krishnan E, Bruce B. Progress in assessing physical function in arthritis: PROMIS short forms and computerized adaptive testing. Journal of Rheumatology. 2009. 36(9):2061-6
9. Segawa E, Schalet B, Cella D. A comparison of computer adaptive tests (CATs) and short forms in terms of accuracy and number of items administrated using PROMIS profile. Qual Life Res [Internet]. 2020 Jan 8;29(1):213–21. Available from: http://link.springer.com/10.1007/s11136-019-02312-8
10. Czerwonka N, Gupta P, Desai SS, Hickernell TR, Neuwirth AL, Trofa DP. Patient-reported outcomes measurement information system instruments in knee arthroplasty patients: a systematic review of the literature. Knee Surg Relat Res. 2023 Dec;35(1):27.
11. Stephan A, Stadelmann VA, Preiss S, Impellizzeri FM. Measurement properties of PROMIS short forms for pain and function in patients receiving knee arthroplasty. J Patient-Reported Outcomes. 2023 Feb;7(1):18.
12. Dhollander O, Roorda LD, Diarra S, Ghijselings I, Demurie A, Terwee CB, et al. Psychometric Properties and Feasibility of PROMIS Computerized Adaptive Tests Compared with Disease-Specific Measures in Knee Arthroplasty. J Bone Jt Surg [Internet]. 2025 107(21):2371-2388 Available from: https://journals.lww.com/ 10.2106/JBJS.24.01348
13. Prinsen CAC, Mokkink LB, Bouter LM, Alonso J, Patrick DL, de Vet HCW, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res. 2018; 27(5):1147-1157
14. Crins MHP, van der Wees PJ, Klausch T, van Dulmen SA, Roorda LD, Terwee CB. Psychometric properties of the PROMIS Physical Function item bank in patients receiving physical therapy. PLoS One. 2018;13(2).
15. Crins MHP, Terwee CB, Klausch T, Smits N, de Vet HCW, Westhovens R, et al. The Dutch–Flemish PROMIS Physical Function item bank exhibited strong psychometric properties in patients with chronic pain. J Clin Epidemiol. 2017;87.
16. Amtmann D, Cook KF, Jensen MP, Chen WH, Choi S, Revicki D, et al. Development of a PROMIS item bank to measure pain interference. Pain. 2010;150(1):173–82.
17. Rose M, Bjorner JB, Gandek B, Bruce B, Fries JF, Ware JE. The PROMIS Physical Function item bank was calibrated to a standardized metric and shown to improve measurement efficiency. J Clin Epidemiol. 2014; 67(5):516-26
18. Terwee CB, Coopmans C, Peter WF, Roorda LD, Poolman RW, Scholtes VAB, et al. Development and validation of the computer-administered animated activity questionnaire to measure physical functioning of patients with hip or knee osteoarthritis. Phys Ther [Internet]. 2014 Feb;94(2):251–61. Available from: http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L372744384

19. Crins MHP, Roorda LD, Smits N, De Vet HCW, Westhovens R, Cella D, et al. Calibration and validation of the Dutch-Flemish PROMIS pain interference Item Bank in patients with chronic pain. PLoS One. 2015;10(7).
20. Hays RD, Bjorner JB, Revicki DA, Spritzer KL, Cella D. Development of physical and mental health summary scores from the patient-reported outcomes measurement information system (PROMIS) global items. Qual Life Res. 2009;18(7).
21. Pellicciari L, Chiarotto A, Giusti E, Crins MHP, Roorda LD, Terwee CB. Psychometric properties of the patient-reported outcomes measurement information system scale v1.2: global health (PROMIS-GH) in a Dutch general population. Health Qual Life Outcomes. 2021;19(1).
22. Roos EM, Toksvig-Larsen S, E.M. R, S. T-L, Roos EM, Toksvig-Larsen S. Knee injury and Osteoarthritis Outcome Score (KOOS) - Validation and comparison to the WOMAC in total knee replacement. Health Qual Life Outcomes [Internet]. 2003 May;1:17. Available from: http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L39111979
23. Perruccio A V., Stefan Lohmander L, Canizares M, Tennant A, Hawker GA, Conaghan PG, et al. The development of a short measure of physical function for knee OA KOOS-Physical Function Shortform (KOOS-PS) - an OARSI/OMERACT initiative. Osteoarthr Cartil [Internet]. 2008 May;16(5):542–50. Available from: http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L50072174
24. 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;15(12):1833–40.
25. Dawson J, Fitzpatrick R, Carr A, Murray D. Questionnaire on the perceptions of patients about total hip replacement. J Bone Jt Surg - Ser B. 1996;78(2):185–90.
26. De Vet HCW, Terwee CB, Mokkink LB, Knol DL. Measurement in medicine: A practical guide. Measurement in Medicine: A Practical Guide. 2011.
27. 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;63(7):737–45.
28. Braaksma C, Wolterbeek N, Veen MR, Poolman RW, Pronk Y, Klaassen AD, et al. Assessing the measurement properties of PROMIS Computer Adaptive Tests, short forms and legacy patient reported outcome measures in patients undergoing total hip arthroplasty. J Patient-Reported Outcomes. 2024 Oct;8(1):121.
29. Kendall R, Wagner B, Brodke D, Bounsanga J, Voss M, Gu Y, et al. The relationship of PROMIS pain interference and physical function scales. Pain Med (United States). 2018;19(9).
30. Nixon DC, McCormick JJ, Johnson JE, Klein SE. PROMIS Pain Interference and Physical Function Scores Correlate With the Foot and Ankle Ability Measure (FAAM) in Patients With Hallux Valgus. Clin Orthop Relat Res. 2017;475(11).

PART III

Standardizing legacy PROM score conversions towards PROMIS scores

CHAPTER 6

Validating existing crosswalks between PROMs and PROMIS measuring physical functioning in patients undergoing total hip and total knee arthroplasty

Advances in Patient-Reported Outcomes. 2025 Oct 1; in press.

Abstract

Objectives

Several crosswalks are available for frequently used legacy patient-reported outcome measures (PROMs) in orthopedic surgery evaluating outcomes after total hip arthroplasty (THA) or total knee arthroplasty (TKA). This multicenter study validates crosswalks of the Hip disability and Osteoarthritis Outcome Score Physical function Shortform (HOOS-PS), the Knee injury and Osteoarthritis Outcome Score -Physical function Shortform (KOOS-PS), and the KOOS Function in daily living (ADL) subscale to Patient-Reported Outcomes Measurement Information System (PROMIS) SF10a physical function (PROMIS PF) scores.

Methods

Patients of three orthopedic departments completed online questionnaires: HOOS-PS, KOOS-PS, KOOS ADL and PROMIS PF. After converting the legacy PROM scores towards PROMIS metric using the crosswalk tables, the Pearson's correlation and the intraclass correlation coefficient (ICC) between the predicted and observed PROMIS scores were calculated. The level of agreement between the predicted and observed PROMIS scores was assessed using a Bland-Altman plot, and the Limits of Agreement (LoA) were calculated. Last, the percentage of patients for whom the predicted score was considered acceptably comparable to the observed score (difference ≤2 points) was determined.

Results

422 patients were included. Adequate correlations (≥0.70) and ICC values (≥0.70) were found between observed and predicted PROMIS scores, indicating good performance of the crosswalks and suggesting good agreement on group level. Mean differences between predicted and observed PROMIS scores based on the HOOS-PS, KOOS-PS, and KOOS ADL were respectively 1.4, -0.3 and 1.0. The LoA varied between -10.8 and 12.4, indicating substantial differences between observed and predicted PROMIS score on individual patient level. Only 25.7%-39.8% of the patients had a predicted PROMIS score acceptably comparable to the observed PROMIS score.

Conclusion

The existing HOOS-PS, KOOS-PS and KOOS-ADL crosswalks towards PROMIS PF seem to be appropriate for group-level use but are not suitable for individual-level predictions of PROMIS scores.

Introduction

The Patient-Reported Outcomes Measurement Information System (PROMIS® (1,2)) is a collection of high-quality patient-reported outcome measures (PROMs) that are increasingly used in orthopedic practice(2,3). PROMIS is psychometrically based on Item Response Theory (IRT), which models the relationship between an individual's performance on a test and the underlying traits of that individual (e.g., the level of pain or functional limitations(4). IRT item banks comprise of large sets of questions, ordered by difficulty along an underlying metric. PROMIS measures are scored on a common metric and are developed to measure general health domains, including physical function, pain, fatigue, and emotional well-being. These domains are common across many conditions. Therefore, PROMIS can be used in a disease-transcending manner. Consequently, this promotes more uniformity in data collection across different populations, enabling its use across conditions and for patients with multimorbidity (1,5). PROMIS can be used with short forms or Computer Adaptive Testing (CAT).

The conversion of scores of existing traditional ("legacy") PROMs into PROMIS scores is crucial for maintaining continuity in longitudinal studies and registries when one transitions from using legacy PROMs to PROMIS, enabling comparisons across time. Furthermore, converting outcome scores of existing literature enables comparisons to future studies. The transition of these legacy PROMs to PROMIS presents a notable challenge. Algorithms for converting historical data into the new PROMIS scores are needed. For this purpose, crosswalks have been developed. These crosswalks enable score conversions between legacy PROMs and PROMIS. Crosswalks are mapping algorithms that convert scores from one instrument to another, allowing consistency in interpreting results across different measurement tools and facilitating comparability and continuity of data in clinical practice and research (Schalet et al., 2021).

Frequently used legacy PROMs in orthopedic surgery evaluating outcomes after THA or TKA are the Hip disability and Osteoarthritis Outcome Score Physical function Shortform (HOOS-PS (6)), the Knee injury and Osteoarthritis Outcome Score -Physical function Shortform (KOOS-PS, (7)), and the KOOS Function in daily living (ADL) subscale (8). Since the transitioning towards PROMIS, crosswalks have been made to link these legacy PROMs towards the PROMIS Physical Function (PROMIS PF(9,10)) (11–13). These crosswalk studies are often based on multicenter data obtained in the USA. However, these crosswalks are not externally validated. The results may vary across different countries and health conditions.

Therefore, this study seeks to validate the existing crosswalks of the HOOS-PS, KOOS-ADL, and KOOS-PS in patients awaiting THA and TKA or patients who have undergone THA and TKA. With this study, we aim to externally validate the crosswalks from these legacy instruments to PROMIS instruments in terms of reliability and agreement in a multicenter study outside the USA. Furthermore, external validation will prevent the creation of multiple crosswalks for a single legacy instrument, reducing a potential source of confusion and supporting standardization. This research will contribute to standardizing PROM score conversions, facilitating consistent interpretation across clinical settings, and ensuring data continuity across historical and future outcome assessments.

Methods

This study is a prospective, multicenter cohort study. Three high-volume orthopedic departments in the Netherlands included patients undergoing THA and TKA: St. Antonius Hospital in Utrecht, Kliniek ViaSana in Mill, and OLVG in Amsterdam. The study was conducted in accordance with the principles of the Declaration of Helsinki. The study was reviewed by a Medical Ethics Review Committee (MEC-U) (St. Antonius Hospital, Nieuwegein, the Netherlands) (W21.037), which confirmed that the Medical Research Involving Human Subjects Act (WMO) does not apply. With this waiver, approval was obtained from the Institutional Review Board of each participating center. Statistical analysis was performed using IBM SPSS (version 29).

Study participants

Two cohorts were included: (1) patients awaiting THA or TKA surgery and (2) patients who had undergone THA or TKA. Patients in the second cohort were included at 3, 6, or 12 months postoperatively. From the study start date, consecutive patients placed on the waiting list prior to surgery, as well as those presenting at three, six or twelve months postoperatively, were included in the study. Therefore, each patient completed a single test-retest assessment. To qualify for inclusion, patients must be 18 or older. The exclusion criteria were patients who had undergone THA due to a femoral neck fracture and those unable to complete Dutch questionnaires independently. The informed consent form was signed electronically.

Procedures

Each hospital was required to include at least 50 patients at different measurement points (pre and postoperatively). Each patient filled out the questionnaire at a single measurement point. Patients completed an online questionnaire for routine outcome measurement using a web-based platform (OnlinePROMS, Interactive Studios, 's-Hertogenbosch, the Netherlands). This certified (ISO27001; NEN7510) online PROMs platform is linked to the Dutch-Flemish Assessment Center CAT software. When the patient did not respond, a maximum of two automatic reminders were sent every two days after the first invitation. As the digital questionnaire restricts the skipping of items, no item responses were missing.

Measures

Questionnaires were digitally collected and consist of patient characteristics (sex, age, indication and date of surgery), legacy PROMs and PROMIS short forms. The legacy PROMs are part of routine outcome measurement, as advised by the Netherlands Orthopaedic Association (NOV) and results are collected in the Dutch Arthroplasty Register (LROI). This study will only present the (data of the) legacy measurement instruments for which a crosswalk is available towards PROMIS PF. Table 6.1 summarizes the existing crosswalks. This included for patients undergoing THA the HOOS-Physical function Shortform (HOOS-PS (6)), and for patients undergoing TKA the Knee Injury and Osteoarthritis Outcome Score, Function in daily living (KOOS-ADL(8)), and KOOS- Physical Function Shortform (KOOS-PS (7)). Furthermore, the Dutch-Flemish PROMIS v2.0 short form measuring Physical Function 10a (PROMIS PF SF 10a, 10 items) was completed. PROMIS measurement instruments use a T-score metric with a mean of 50 and a standard deviation of 10, where a score of 50 represents the general population average(14). Higher T-scores indicate greater levels of the measured concept (i.e., better function or more pain). The HOOS-PS and KOOS-PS are short versions of the HOOS and KOOS (respectively 5 and 7 items) and the KOOS-ADL (17 items) is a subscale of the KOOS. These legacy measures are reported on a scale of 0 to 100 (0 indicating extreme symptoms), whereby the total score reflects the sum of the individual's responses to each item.

Statistical analysis

Means, standard deviations, and range were calculated for the patient characteristics and their scores on outcome measures. The scores on legacy PROMs were converted into the PROMIS metric using the respective crosswalks (Table 6.1). These scores were

defined as predicted PROMIS scores. The observed scores were the measured PROMIS scores.

Table 6.1. Existing crosswalks between legacy instruments and PROMIS measures.

Legacy instrument	PROMIS instrument	Crosswalk table
HOOS-PS[6]	PROMIS Physical Function Short form 10a[9,10]	HOOS-PS+table.pdf (squarespace.com)[12]
KOOS-PS[7]	PROMIS Physical Function Short form 10a[9,10]	KOOS-PS+table.pdf[13]
KOOS-ADL[8]	PROMIS Physical Function Short form 10a[9,10]	KOOS-ADL+table.pdf (squarespace.com)[11]

To compare the observed and predicted PROMIS scores, the following methods were used:

1. Correlation

Pearson's correlations between the legacy instruments and observed PROMIS scores were calculated to assess if the PROMs sufficiently assess the same construct. The correlations between the HOOS-PS, KOOS-PS, and KOOS ADL, with PROMIS PF were expected to be high since they aim to measure the same construct (physical function). In addition, the correlations between the predicted and observed PROMIS scores were calculated to evaluate the performance of the crosswalks. A good correlation was defined as a minimum of Pearson $r = 0.70$[15].

2. Intraclass correlation coefficient (ICC)

To estimate the agreement between predicted and observed PROMIS scores the ICC was calculated. The ICC was calculated using a two-way random-effects model for absolute agreement: $ICC\ agreement = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_m^2 + \sigma_e^2}$, whereby σ_p^2 is the variation between patients, σ_m^2 is the systematic variation between observed and predicted scores, and σ_e^2 is random error variance. An ICC ≥0.70 was considered adequate for group-level comparisons[16].

3. Standardized mean difference (Cohen's d)

To assess the level of agreement on the group level, standardized mean differences (SMD/Cohen's d) for paired samples were calculated, along with 95% confidence intervals (CI). The SMD was calculated with the following formula: $SMD = \frac{(\mu_B - \mu_A)}{SDd}$, whereby *SDd* is the standard deviation of difference scores, μ_A is the mean of group A and μ_B is the mean of group B. Previous linking validation studies established a 0.2 threshold for the SMD (also referred to as Cohen's d), which is comparable to 2 T-score points. An SMD value of less than 0.2 was considered negligible[17–19].

4. Bland-Altman Plot

In addition to the ICC, the level of agreement between the predicted and observed PROMIS scores was further assessed using a Bland-Altman plot. This method visually presents the mean difference between the predicted and observed PROMIS scores against their average values. Limits of agreement (LoA) were calculated (mean difference ± 1.96*standard deviation). The LoA indicate the range where 95% of the difference between the observed and predicted scores lie.

5. The adequacy of the predicted PROMIS scores for the individual patient

The percentage of patients for whom the predicted score was considered acceptably comparable to the observed score were determined. The predicted score was considered acceptably comparable if the difference between the predicted PROMIS and observed PROMIS scores was ≤2 points, as a difference of 2 points has been found meaningful to patients[20].

Results

The multicenter study included 422 patients with a mean age of 67.9 years (SD 8.5). Compared to the study populations used to develop the existing crosswalks, the sample of this study is comparable regarding age, sex, and indication. The patient demographics of the study samples and scores on outcome measures are presented in Table 6.2.

Correlation

The Pearson's correlation coefficients between the scores on legacy measures and PROMIS measures are presented in Table 3. Correlations ≥0.70 were found between the HOOS-PS, KOOS-PS, and KOOS ADL with PROMIS PF SF10a. This indicates that these instruments sufficiently assess the same construct: physical function. Table 4 presents the Pearson's r correlations between the predicted PROMIS scores and observed PROMIS scores. Correlations ≥0.70 were found between observed and predicted PROMIS SF10a scores, indicating good performance of the crosswalks.

Table 6.2. Patient demographics and summary statistics of the study samples.

Variables	**Braaksma TKA**	**Braaksma THA**	**Tang** (13)	**Heng** (12)	**Heng** (11)
N	216	206	3667	3382	1003
Age (mean, (SD))	68.4(7.7)	67.3(9.2)	66.4(10.7)	66.4(10.7)	67(8)
Sex – female(%)	48.6%	61.7%	56.9%	56.9%	60%
Health condition	TKA preoperative 56%, 6 months postop 31%, 12 months postop 13%	THA preoperative 56%, 3 months postop 18%, 6 months postop 11%, 12 months postop 13%	Consideration of a primary TKA non-surgical (n= 933), preoperative, postoperative	Hip osteoarthritis (n=2097), preoperative, postoperative	98% knee osteoarthritis
Scores (mean (SD); range)					
HOOS-PS	*	68.9(16); 17.6-100	*	71.9(19.8); 0-100	*
KOOS-PS	40 (17); 0-100	*	*	*	*
KOOS-ADL	65.4 (23.4); 6-100	*	*	*	68(22)
PROMIS PF SF10a	39 (9) 20.9-61.9	39.1 (7) 20.9-61.9	*	41.3(9), 13.5-61.9	41(8)
Predicted PROMIS PF SF 10a from HOOS-PS	*	40.5 (10.1) 23-59.4	*	*	*
Predicted PROMIS PF SF 10a from KOOS-PS	38.7(7.1) 19-60.9	*	*	*	*
Predicted PROMIS PF SF 10a KOOS-ADL	39.9(7.8) 22.8-60.8	*	*	*	*

Abbreviations: TKA = total knee arthroplasty, THA = total hip arthroplasty, SD = standard deviation.
* : no data available

Table 6.3. Correlation coefficient between legacy measures and PROMIS measure.

Sample	**Legacy measure**	**PROMIS measure**	**Pearson's (r)**
THA (n=206)	HOOS-PS	PROMIS PF SF10a	-0.81
TKA (n=216)	KOOS-PS	PROMIS PF SF10a	-.71
	KOOS-ADL/function	PROMIS PF SF10a	-.78

*HOOS-PS and KOOS-PS show negative correlation due to the fact that the direction of scores are reversed.

Interclass correlation coefficient

The ICC assessed both the correlation and consistency between the observed and predicted PROMIS scores. ICC values ≥0.70 were found, suggesting adequate agreement between observed and predicted scores on group level (Table 6.4).

Standardized mean difference / Cohen's d

SMDs of <0.2 were found for the paired differences in the TKA group, indicating negligible effect sizes for any observed differences (Table 6.4). The SMD of the paired

differences between the predicted PROMIS PF from the HOOS-PS and the observed PROMIS PF was slightly higher (.25, Table 6.4).

Table 6.4. Correlation coefficient and the Interclass correlation coefficient (ICC) between the predicted PROMIS scores and observed PROMIS scores.

Sample	Predicted scores of the legacy instruments towards PROMIS scores	Observed PROMIS scores	Pearson's (r)	ICC, single measures	Mean difference (points, SD)	SMD / Cohen's d (SD differences) (95% CI)
THA (n=206)	Predicted PROMIS PF from HOOS-PS	PROMIS PF SF10a	.83	.82	1.4(5.6)	-.25(-.39; -.14)
TKA (n=216)	Predicted PROMIS PF from KOOS-PS	PROMIS PF SF 10a	.71	.71	-0.3(5.4)	-.06(-.19; -.08)
	Predicted PROMIS PF from KOOS-ADL/function	PROMIS PF SF 10a	.77	.76	1.0(5.1)	.19(.06; -.33)

Bland - Altman plot

The discrepancies between predicted and observed scores at each score point were graphically shown using Bland-Altman plots (Figure 6.1). Mean differences between predicted and observed PROMIS PF SF10a scores based on the HOOS-PS, KOOS-PS, and KOOS ADL were respectively 1.4, -0.3 and 1.0. The Level of Agreement (LoA) obtained from the Bland & Altman analysis show the range within 95% of the differences between observed and predicted scores fall and are plotted as red horizontal lines in Figure 6.1. The lower and upper LoA were respectively -9.6 - 12.4 for the HOOS-PS, -10.8 - 10.2 KOOS-PS and -9 - 11 KOOS-ADL. This indicates substantial differences between observed and predicted PROMIS scores on individual patient level.

The adequacy of the predicted PROMIS scores for the individual patient

The percentage of patients with a predicted PROMIS SF10a score acceptably comparable to the observed PROMIS score (≤2 points difference between the observed score and predicted score) was 25.7% on the HOOS-PS, 30.6% on the KOOS-PS, and 39.8% on the KOOS-ADL.

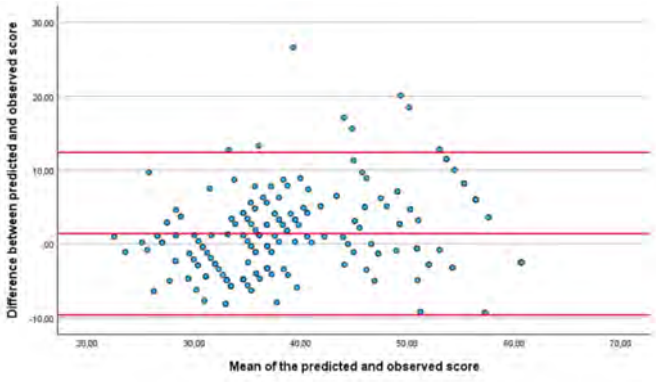

Figure 6.1a. The LoA between the observed and predicted PROMIS SF10a scores based on the HOOS-PS (THA sample, n=206).

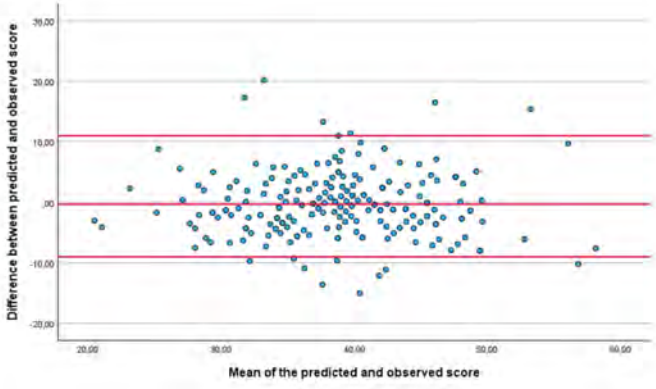

Figure 6.1b. The LoA between the observed and predicted PROMIS SF10a scores based on the KOOS-PS (TKA sample, n=216).

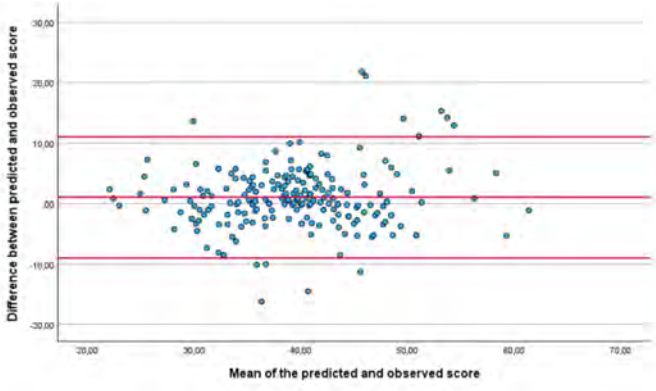

Figure 6.1c. The LoA between the observed and predicted PROMIS PF SF10a scores based on the KOOS-ADL (TKA sample, n=216).

Discussion

In this study, we externally validated existing crosswalks for predicting PROMIS PF scores from legacy instruments (HOOS-PS, KOOS-PS, and KOOS-ADL) in patients undergoing THA or TKA. We found that the existing crosswalks are appropriate for group-level use but are not suitable for individual-level predictions of PROMIS scores.

Our findings confirm the general thought that crosswalks are not suitable for individual patient-level interpretation and analysis. Only 25-40% of the predicted scores were considered acceptably comparable to the observed scores, underscoring the significant discrepancies between predicted and observed scores at the individual level, a concern that has been extensively documented[21]. However, as Hartman et al. note, these discrepancies may be attributable to the unreliability of the PROM itself, rather than the linking procedure, as evident from the limits of agreement of PROM test-retest analyses[22]. Nevertheless, the limited predictive accuracy highlights the need for caution when using predicted PROMIS scores in clinical evaluations of individual patients.

Consistent with the findings of the development studies, the agreement at the group level for the HOOS-PS, KOOS-PS, and KOOS-ADL crosswalks was satisfactory, with predicted scores comparable to the observed scores. Therefore, group-level analysis using these crosswalks seems feasible, although this feasibility may depend on factors such as patient selection and sample size.

Though the mean differences between observed and predicted PROMIS scores were comparable, the SMD for HOOS-PS was 0.25, slightly exceeding the cut-off of 0.2. The discrepancy can be attributed to several factors. First, the optimal methodology for crosswalk development remains underexplored, with variations in linking methods potentially yielding different crosswalks[17,24]. Second, the limited validity of the HOOS-PS, including substantial measurement error, likely contribute to the reduced accuracy of the resulting predicted scores[25]. Furthermore, the SMD varies depending on the method used. In particular, using a pooled SD typically rather leads to a smaller SMD than using the SD of the difference score (used in this study since it is a paired sample).

Finally, the external validation conducted in this study might account for some of the differences. However, the sample in this study seemed similar in terms of age, sex, and patient selection with those of Tang et al., and the correlations between predicted and observed scores were comparable (respectively .70 and .78 [13]).

Our literature search shows that there is currently only one crosswalk available per measurement instrument. The development of multiple crosswalks for a single instrument could lead to data inconsistency. While only a few previous studies have externally validated crosswalks from disease-specific to general PROMs, they focused on different healthcare domains such as lower back pain and depressio [22,26]. Similar to our findings, these studies demonstrated that the prediction of PROMIS scores using crosswalks from legacy scores was comparable at group level.

This study has several strengths. The total sample size included 422 patients across multiple participating sites, enhances the reliability of group-level agreements. Although no formal a priori sample size calculation was performed, this sample size aligns with published guidelines recommending a minimum of 150 participants to estimate an ICC of 0.70 with a 95% confidence interval width of ±0.10[27]. Our sample, therefore, provides sufficient support for group-level interpretations of agreement between predicted and observed PROMIS scores. Previous studies have indicated that crosswalks can yield reliable results, as shown by Choi et al. 2014. The standard error of the predicted T-score was minimized using just 75 patients[28]. Another methodological strength is the validation of crosswalks using independent data from another country and language in this study. Therefore, the validation is a true test of the performance of the existing crosswalks. Additionally, the multi-center study design and the use of different versions of Dutch-Flemish questionnaires further contributed to the quality of this study.

Nonetheless, several limitations of this study should be considered. We did not include patient characteristics such as race or socioeconomic status, although these factors can influence the responses. Furthermore, the chosen cut-off MIC value for acceptably comparability of predicted and observed PROMIS scores in this study may influence group-level agreement. The acceptable difference was set on two T-point scores in this study, based on Terwee et al. (MIC 2-6[20]), while another study used a MIC of three points T-score as threshold[26]. The conservative threshold was adapted to ensure that potential differences in score equivalence were not underestimated. A lower threshold leads to a greater applicability of the crosswalk. In patients undergoing arthroplasty, the expected improvement is approximately 11 PROMIS T-score points (IQR 6-17)[23]. However, when applying crosswalks in comparing baseline scores of populations or when expecting minor improvements, it is important to have this conservative threshold. Other choices of an acceptable difference between predicted and observed scores are justifiable. For example, an alternative for comparing predicted versus observed scores is reporting the percentage of patients where the predicted score fell within the confidence intervals (e.g., ±2×SE) of the observed PROMIS score.

Future research should focus on developing crosswalks for widely used legacy PROMs, such as the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), Oxford Knee Score (OKS), and Oxford Hip Score (OHS), which currently lack crosswalks. This represents a significant gap in the literature that remains unaddressed. When developing new crosswalks, it is important to conduct external validation to confirm the applicability of the crosswalks in various settings. External validation is a crucial step before implementing crosswalks in clinical practice.

Conclusion

In conclusion, the existing HOOS-PS, KOOS-PS, and KOOS-ADL crosswalks for predicting PROMIS PF seem to be suitable for group-level use. However, these crosswalks should not be utilized for individual-level predictions of PROMIS scores.

Impact statement

This research contributes to standardizing PROM score conversions, facilitating consistent interpretation across clinical settings, and ensuring data continuity across historical and future outcome assessments in patients undergoing hip or knee arthroplasty.

Credit authorship contribution statement

Y. Pronk: Writing – review & editing, Validation, Project administration, Methodology, Investigation, Data curation. **A.D. Klaassen:** Project administration, Methodology, Investigation, Data curation. **R.W. Poolman:** Writing – review & editing, Validation, Supervision. **D. Delawi:** Writing – review & editing, Supervision, Data curation. **B.D. Schalet:** Writing – review & editing, Validation, Supervision, Methodology, Formal analysis. **R.W.J.G. Ostelo:** Writing – review & editing, Validation, Supervision, Methodology, Conceptualization. **C.B. Terwee:** Writing – review & editing, Validation, Supervision, Methodology, Conceptualization. **Braaksma Christel:** Writing – original draft, Visualization, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **B.G.N. Mobers:** Writing – original draft, Project administration, Formal analysis. **L.W.A.H. van Beers:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of interests

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Schalet B reports a relationship with PROMIS Health Organization that includes: board membership. Schalet B reports a relationship with Dutch-Flemish PROMIS National Center that includes: board membership. Schalet B reports a relationship with Northwestern University that includes: funding grants. C.B. Terwee: PROMIS Health Organization, Chair International Committee. C.B. Terwee: Given her role as Editor- in-Chief, C.B.T had no involvement in the peer-review of this article and has no access to information regarding its peer-review. Full responsibility for the editorial process for this article was delegated to another journal editor. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

None

Funding sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The data of this study is used from a previously granted study which was funded by the LROI (LROI RG 2021–001). The LROI had no role in data analysis or interpretation.

Ethics approval

The study was conducted in accordance with the Declaration of Helsinki, and approved by the Medical Ethics Review Committee (MEC- U) in the Netherlands, which confirmed that the Medical Research Involving Human Subjects Act (WMO) does not apply (registration code: W21.037 and date of approval: 18 February 2021). With this waiver, approval of the Institutional Review Board of each participating center was obtained.

References

1. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. The patient-reported outcomes measurement information system (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. J Clin Epidemiol. 2010;63(11):1179–94.
2. Horn ME, Reinke EK, Couce LJ, Reeve BB, Ledbetter L, George SZ. Reporting and utilization of Patient-Reported Outcomes Measurement Information System® (PROMIS®) measures in orthopedic research and practice: a systematic review. Journal of Orthopaedic Surgery and Research. 2020; 15(1):553.
3. Wong LH, Meeker JE. The promise of computer adaptive testing in collection of orthopaedic outcomes: an evaluation of PROMIS utilization. Journal of Patient-Reported Outcomes. 2022; 6(1):2.
4. De Ayala RJ. The theory and practice of item response theory. Guilford Press; 2009. 643 p.
5. Terwee C, Ahmed S, Alhasani R, Alonso J, Bartlett S, Chaplin J, et al. Comparable real-world patient-reported outcomes data across health conditions, settings, and countries: the PROMIS international collaboration. NEJM Catal. 2024;5(9).
6. Davis AM, Perruccio A V., 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. Osteoarthr Cartil [Internet]. 2008 May;16(5):551–9. Available from: http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L50072174
7. Perruccio A V., Stefan Lohmander L, Canizares M, Tennant A, Hawker GA, Conaghan PG, et al. The development of a short measure of physical function for knee OA KOOS-Physical Function Shortform (KOOS-PS) - an OARSI/OMERACT initiative. Osteoarthr Cartil [Internet]. 2008 May;16(5):542–50. Available from: http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L50072174
8. Roos EM, Roos HP, Lohmander LS, Ekdahl C, Beynnon BD. Knee Injury and Osteoarthritis Outcome Score (KOOS) - Development of a self-administered outcome measure. J Orthop Sports Phys Ther. 1998;28(2):88-96
9. Crins MHP, Terwee CB, Klausch T, Smits N, de Vet HCW, Westhovens R, et al. The Dutch–Flemish PROMIS Physical Function item bank exhibited strong psychometric properties in patients with chronic pain. J Clin Epidemiol. 2017;87:47-58.
10. Crins MHP, van der Wees PJ, Klausch T, van Dulmen SA, Roorda LD, Terwee CB. Psychometric properties of the PROMIS Physical Function item bank in patients receiving physical therapy. PLoS One. 2018;13(2).
11. Heng M, Tang X, Schalet BD, Collins AK, Chen AF, Melnic CM, et al. Can the Knee Outcome and Osteoarthritis Score (KOOS) function subscale be linked to the PROMIS physical function to crosswalk equivalent scores? Clin Orthop Relat Res. 2021;479(12).
12. Heng M, Stern BZ, Tang X, Schalet BD, Collins AK, Chen AF, et al. Linking Hip Disability and Osteoarthritis Outcome Score-Physical Function Short Form and PROMIS Physical Function. J Am Acad Orthop Surg. 2022;30(15).
13. Tang X, Schalet BD, Heng M, Lange JK, Bedair HS, O'Brien TM, et al. Linking the KOOS-PS to PROMIS Physical Function in Knee Patients Evaluated for Surgery. J Am Acad Orthop Surg. 2022;30(6).
14. Terwee CB, Roorda LD. Country-specific reference values for PROMIS® pain, physical function and participation measures compared to US reference values. Ann Med. 2023;55(1).
15. Cella D, Schalet B, Kallen M, Lai J-S, Cook K, Rutsohn J, et al. PROSETTA stone analysis report: A rosetta stone for patient reported outcomes. 2016.
16. Lohr KN. Assessing health status and quality-of-life instruments: Attributes and review criteria. Quality of Life Research. 2002;11(3):193-205.
17. Schalet BD, Lim S, Cella D, Choi SW. Linking Scores with Patient-Reported Health Outcome Instruments:A VALIDATION STUDY AND COMPARISON OF THREE LINKING METHODS. Psychometrika. 2021;86(3).
18. Cohen J. Statistical Power Analysis for the Behavioral Sciences. Routledge; 2013.
19. Brinker AY, Nolte S, Fischer FH, Obbarius A, Rose M, Liegl G. Comparing Approaches to Link SF-36 PF-10 Scores to PROMIS Physical Function: A Validation Study in Three Clinical Samples. J Gen Intern Med. 2025;40(14):3326-3334
20. Terwee CB, Peipert JD, Chapman R, Lai JS, Terluin B, Cella D, et al. Minimal important change (MIC): a conceptual clarification and systematic review of MIC estimates of PROMIS measures. Vol. 30, Quality of Life Research. 2021;30(10):2729-2754.

21. Ackerman IN, Soh SE, Hallstrom BR, Fang Y., Franklin PD, Lutzner J, et al. A systematic review of crosswalks for converting patient-reported outcome measure scores in hip, knee, and shoulder replacement surgery. Acta Orthop. 2024;13(95):512–23.
22. Hartmann C, Liegl G, Rose M, Fischer F. Towards Standardized Assessment of Outcomes in Back Pain-Validation of Linking Studies Between Disease-Specific and Generic Patient-Reported Outcome Measures. J Clin Med. 2024;13(21).
23. Leal J, Holland CT, Easley ME, Nunley JA, Ryan SP, Bolognesi MP, et al. Comparison of PROMIS scores after total hip and total ankle arthroplasty : a propensity score-matched study. Bone Jt open. 2025 May 1;6(5 Supple A):1–13.
24. Mansolf M, Blackwell CK, Cella D, Lai JS. Assessing the interchangeability of linked scores in multivariable statistical analyses. Qual Life Res. 2024;33(4).
25. Braaksma C, Wolterbeek N, Veen MR, Prinsen CAC, Ostelo RWJG. Systematic review and meta-analysis of measurement properties of the Hip disability and Osteoarthritis Outcome Score - Physical Function Shortform (HOOS-PS) and the Knee Injury and Osteoarthritis Outcome Score - Physical Function Shortform (KOOS-PS).Osteoarthritis and Cartilage. 2020;28. 1525–38.
26. Tang X, Schalet BD, Janulis P, Kipke MD, Kaat A, Mustanski B, et al. Can a linking crosswalk table be applied to a different population? An independent validation study for a crosswalk between BSI depression and PROMIS depression scales. PLoS One. 2022;17(11):e0278232.
27. Mokkink LB, de Vet H, Diemeer S, Eekhout I. Sample size recommendations for studies on reliability and measurement error: an online application based on simulation studies. Heal Serv Outcomes Res Methodol [Internet]. 2023 Sep 23;23(3):241–65. Available from: https://link.springer.com/10.1007/s10742-022-00293-9
28. Choi SW, Schalet B, Cook KF, Cella D. Establishing a common metric for depressive symptoms: Linking the BDI-II, CES-D, and PHQ-9 to PROMIS Depression. Psychol Assess. 2014;26(2).

CHAPTER 7

General discussion

Summary of main results

This thesis examined the psychometric properties of current patient-reported outcome measures (PROMs) assessing physical function and pain in total hip arthroplasty (THA) and total knee arthroplasty (TKA). Furthermore, it has compared the currently used ("legacy") PROMs with a new innovative set of measures, the Patient-Reported Outcomes Measurement Information System (PROMIS®), and explored the external validation of linking data between legacy PROMs and PROMIS instruments. Please note that in this general discussion, the term PROMs refers specifically to legacy PROMs.

The most critical finding of this thesis is the identification of limitations concerning the reliability and validity of the PROMs currently used for THA and TKA patients. Inaccurate measurement burdens patients and health-care deliverers unnecessarily, and reliance on data with limited validity risks misguided clinical decision-making. The innovative alternative, PROMIS, seems more suitable for individual patient assessment, offering a lower burden and a wider measurement range. In case transitioning from legacy PROMs to PROMIS measurement, existing crosswalks from these legacy PROMs to PROMIS scores can be used for group-level data transitions.

As explained in the introduction of this manuscript, the need for the implementation of valid PROM is evident, as both patients and healthcare providers require accurate evaluations that facilitate the use of PROM data in research and inform policy initiatives. However, despite their potential benefits, numerous challenges remain in the selection and implementation of the best PROMs in practice. Not implementing the most adequate PROMs results in less efficient and less patient-centered healthcare.

This thesis is not merely a research endeavor but serves as a call for practice improvement. While its primary focus is methodological, the implications extend to clinical decision-making, postoperative monitoring, and value-based healthcare.

Implementation of PROMs in THA and TKA: facilitators, barriers and recommendations

In the following section, we will employ the Consolidated Framework for Implementation Research (CFIR[1]) to systematically structure and interpret the key challenges encountered during the implementation of PROMs within Dutch Healthcare. CFIR serves as a practical framework to guide systematic assessment of potential

barriers and facilitators for implementation strategies. The framework comprises five domains: Intervention Characteristics, Outer Setting, Inner Setting, Characteristics of Individuals, and Process, each illuminating specific barriers and facilitators. This section will detail the challenges implementing PROMs for THA and TKA, while providing recommendations for solutions.

I. PROM characteristics

Measuring with PROMs brings specific challenges and opportunities. This section will discuss several barriers and facilitators of the PROM characteristics for implementation.

Barrier: invalid PROMs

A common valid metric for measuring outcomes is essential to ensure that we are truly measuring what matters to patients. The PROMs currently recommended by the Dutch Orthopedic Society (NOV) have substantial concerns regarding their measurement properties. As outlined in Chapter 3 of this thesis, the HOOS-PS and KOOS-PS do not adequately reflect physical functioning of THA and TKA patients[2,3]. Consequently, this work has led to the decision that these questionnaires are no longer mandatory in the Netherlands[4]. Nevertheless, the International Consortium for Health Outcomes Measurement (ICHOM) continues to advocate for the use of these questionnaires in evaluating patients with hip and knee osteoarthritis[5]. Consequently, the collected data cannot be optimally used for either research or clinical practice. For instance, physical functioning in patients undergoing THA or TKA is currently measured using PROMs with limitations in validity, a reality encapsulated in the phrase “garbage in, garbage out”, where invalid measures yield less useful data.

This issue highlights a broader problem within orthopedic practice: an abundance of poorly validated but widely used PROMs. Similar to the findings in Chapter 2 and 3, legacy PROMs among THA and TKA patients have limitations in content validity and responsiveness[2,3,6,7]. Furthermore, measurement error is often too large for individual assessment[8]. Moreover, when evaluating THA and TKA patients, it is important to ensure a broad measurement range (i.e. the ability of a PROM to measure across the full spectrum of a health domain, from very poor to excellent health status). Patients prior to arthroplasty often endure significant levels of pain and functional impairment, while those following total hip arthroplasty can experience minimal pain. Using legacy PROMs can lead to a high number of extreme scores, known as floor effects, shown in chapter 4 and 5[9]. As a result, these PROMs are less suitable for assessing physical function of patients undergoing arthroplasty. Consequently, the effect of the treatment on our patients cannot be optimally evaluated. The proliferation of PROMs leads to unaligned

data standards and precludes data to be compared. The underlying reasons for the limitations in many PROMs vary, insufficient content validity is often the main reason. It is crucial that PROMs should be developed and evaluated with the same rigor as other clinical tests, such as blood tests or imaging. Their non-invasive nature does not exempt them from validation.

Recommendation #1: transitioning towards advanced psychometric measurement

If we decide to evaluate patients undergoing THA or TKA using PROMs, we need to reduce the bunch of data to only useful, valid data. Employing a measurement strategy based on advanced psychometric methods, particularly Item Response Theory (IRT) models can enhance precise, adaptive and linear measurement. An IRT model can be used to calibrate a large set of questions ('items') in the same health domain along an underlying metric, according to their difficulty. This underlying theta metric functions as an interval scale. IRT-based item banks can provide short, adaptable, sustainable and universally applicable PROMs with strong measurement reliability and a standardized scale. Calibrated item banks allow for Computer Adaptive Testing (CAT). This approach tailors the questionnaire to each respondent by selecting the most informative items based on prior answers[10]. Therefore, IRT based instruments are more efficient, reducing respondent burden by selecting only relevant items, while maintaining accuracy. In contrast, currently used ('legacy') PROMs are based on Classical Test Theory (CTT). CTT instruments rely on fixed sets of items and assume that all questions contribute equally to the total score. This fixed structure increases burden and limits measurement precision, as all items contribute equally to the total score regardless of their individual relevance or difficulty. Concluding, IRT based PROMs surpass CTT-based instruments by providing more precise, adaptive and linear measurements[11,12].

Barrier: disease-specific PROMs

The currently advised PROMs for THA and TKA patients, such as the Oxford Hip Score and Oxford Knee score, are disease-specific and thus tailored primarily to capture physical function related to the specific joint. This restricts the ability to compare outcomes across different conditions or with the general population. In addition, the use of multiple disease-specific PROMs across patient groups complicates PROM implementation, data harmonization and benchmarking between registries and healthcare systems. Furthermore, patients with multimorbidity often need to multiple disease-specific PROMs that measure the same construct (e.g., physical function or pain), making it difficult to determine which condition is driving the reported outcome.

Recommendation #2: transition to generic measurement

There is increasing evidence that various outcomes, such as pain and physical function, are important for most individuals regardless of their health status. Routinely measuring these outcomes in all patients, facilitated by a generic instrument, offers an alternative to disease-specific measures. Generic instruments could be used disease-transcending and focus on measuring general health domains such as physical function, pain, fatigue, and emotional well-being. These domains are important across many conditions, promoting standardization in data collection across different populations and enhancing use across conditions and for patients with multimorbidity[13,14]. Therefore, it will reduce the amount of data collection. Another benefit is that it enables benchmarking across conditions, institutions and providers. For instance, physical functioning evaluations following THA or following a percutaneous coronary intervention for myocardial infarction can be conducted with the same instrument. Additionally, generic measures supports the strategic allocation of healthcare resources, especially in the doomsday scenario on comparative healthcare effectiveness, ensuring that interventions with the greatest impact receive priority.

Recommendation #3: implementing a new standard outcome measurement set, transition to PROMIS.

The goal for each measurement instrument is that it should be easily interpretable, adequately validated, and implemented for clinical use. Decision-making regarding the selection of PROMs and the specific constructs to be evaluated in THA and TKA necessitates careful consideration. Regarding THA and TKA, the OMERACT initiative outlines the key domains that should be assessed in patient with osteoarthritis, including pain, physical function, quality of life, and the patient's global assessment of target joints[15]. Prioritizing standardized and validated PROMs is essential to ensure accurate and meaningful assessments of patient-reported physical function and pain. By combining the above mentioned recommendations (IRT based and generic measurement), we can yield comparable PROM data across different patient groups and providers using reliable and valid measures. An advanced example of this approach is the Patient-Reported Outcomes Measurement Information System (PROMIS®), which offers a set of measures that enables the evaluation of relevant patient-reported outcomes across various medical conditions, languages, and countries[13]. PROMIS is available in both short forms and Computer Adaptive Testing (CAT) versions. The adoption of PROMIS will address the lack of validity of legacy PROMs in THA and TKA patients, as PROMIS has demonstrated robust measurement properties, including content validity, structural validity, broad measurement range, responsiveness and measurement precision[9,16–20]. A higher precision is required if a PROM plays a role in deciding to

undergo surgery or evaluate treatment effects at individual patient level than for comparing groups. More reliable outcome scores can ensure more accurate individual patient monitoring, improve reliability of study results and can contribute to increase the use of patient reported outcome measures in the consultation room. Furthermore, both PROMIS CAT and PROMIS SF provide a broader measurement range than legacy PROMs, thereby decreasing floor and ceiling effects[9]. When there is a need to accurately measure patients at the end of the scale (e.g. patients following THA with minimal pain), it is recommended to measure with PROMIS CAT[21].

The best alternative instrument would be the PROMIS CAT PF or the PROMIS SF8b for assessing physical function, instead of the currently recommended Oxford Hip Score or Oxford Knee score. This recommendation is supported by the advice from a working group of the Dutch government (the 'Werkgroep Generieke PROMs'), who developed a standard PRO outcome set ("Generic PROM set") for medical specialist care, including PROMIS measures[22,23]. The PROMIS instruments can be used alongside two Numeric Rating scales measuring pain during rest and activity, a global assessment anchor question regarding function improvement post-surgery, and the EQ-5D for economic evaluations. This proposed transition in PROMs measuring physical function will reduce the number of items questioned with 4-8 items, while ensuring that the selected PROMIS instruments are more relevant and capable of validly measuring outcomes, particularly at the extremes of the measurement scales. To maintain data integrity throughout this transition, existing crosswalks can be used to make scores between the legacy PROMs and PROMIS comparable. These crosswalks have previously been developed for the HOOS-PS and KOOS-PS, and are considered valid for group-level use (Chapter 6)[24]. However, it is important to note that crosswalks for transforming scores from the currently used Oxford Hip Score and Oxford Knee score to PROMIS PF have yet to be established.

Table 7.1 presents a proposal for a new standard outcome measurement set in THA and TKA patients. In addition to pain and physical function, other domains as mental health could be assessed. Psychological and psychosocial factors are well known in pain science to drive outcomes (e.g. depression, anxiety, pain catastrophizing are among the strongest psychosocial predictors of chronic postsurgical pain), but these factors are rarely collected in arthroplasty registries as part of standard outcome measurement. The generic PROM set did include PROMIS mental health measures[22,23]. In the near future it could be that every hospital should implement this generic PROM set. This must be taken into account.

Table 7.1. Proposed new standard outcome measurement set in patients undergoing total hip arthroplasty or total knee arthroplasty.

Currently used PROMs	New standard outcome set
THA	**THA**
NRS pain in rest	NRS pain in rest
NRS pain during activity	NRS pain during activity
EQ-5d-5L	EQ-5d-5L
Oxford Hip Score	PROMIS CAT Physical Function
HOOS-PS(optional)	PROMIS SF8b Physical Function (optional, in case PROMIS CAT cannot be implemented)
Global assessment function improvement (post-surgery)	Global assessment function improvement (post-surgery)
TKA	**TKA**
NRS pain in rest	NRS pain in rest
NRS pain during activity	NRS pain during activity
NRS satisfaction (post-surgery)	NRS satisfaction (post-surgery)
EQ-5d-5L	EQ-5d-5L
Oxford Knee Score	PROMIS CAT Physical Function
KOOS-PS(optional)	PROMIS SF8b Physical Function (optional, in case PROMIS CAT cannot be implemented)
Global assessment function improvement (post-surgery)	Global assessment function improvement (post-surgery)
Global assessment pain reduction (post-surgery)	Global assessment pain reduction (post-surgery)

Recommendation #4: future research

Prior to implementation, it would be beneficial to establish crosswalks between all existing legacy PROMs and PROMIS scores, such as the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), Oxford Knee Score (OKS), and Oxford Hip Score (OHS). Converting scores from legacy PROMs into PROMIS scores will help maintain continuity in longitudinal studies and registries, enabling comparisons across time. Furthermore, minimally important change (MIC) estimates are essential to determine whether changes in scores reflect meaningful changes following THA or TKA. Notably, to date, only one study has addressed the MIC of a PROMIS instrument (PROMIS PF SF) in THA[25], and none in TKA.

Recommendation #5: standards and criteria for good PROMs

In instances where disease-specific PROMs are used , guidelines such as from the COnsensus-based Standards for selecting health status Measurement INstruments (COSMIN) initiative can aid in identifying and developing psychometrically sound measures[26,27]. These guidelines provide standards and criteria for PROM development, validation, and systematic reviews of PROMs. Ideally, medical journal reviewers should

require adherence to these guidelines for PROM development and psychometric evaluation studies.

II. Inner setting – the hospitals in which the PROM should be implemented

Barrier: inadequate integration

For PROMs to be effectively integrated as routine outcome measurement in hospitals, organizational support and resource availability are essential. A consistent barrier to adoption is inadequate IT support, as PROM platforms are often poorly integrated with electronic health records. Last, using PROMs involves costs. These obstacles must be addressed to maximize the utility of PROMs in clinical and research settings[28].

Barrier: inadequate coordination of implementation

Furthermore, a lack of dedicated PROM coordinators or leaders is a frequently reported implementation barrier. Effective institutional leadership that allocates appropriate time and resources can enhance the likelihood of successful implementation, particularly by embedding PROMs into routine workflows and providing clinician training to shift perceptions.

Facilitator: adaption across healthcare organizations and digital infrastructures

A key facilitator identified in the implementation of PROMs is demonstrated through the existing adoption of the PROMIS measures recommended in "`Generic PROM set" across several healthcare organizations and digital infrastructures. The generic PROM set has now been implemented in 35 healthcare institutions and integrated into 15 digital platforms, including KLIK, QuestManager, and OnlinePROMs. Furthermore, integration with commonly used medical record systems has facilitated greater acceptance and sustained engagement among healthcare professionals and institutions. These are critical facilitators for the scalable and sustainable implementation of PROMs within the Dutch healthcare context.

Barrier: data collection and management

Another challenge in the routine PROM measurement, is the data management and data collection. Initiated by the NOV, a nationwide PROM data collection on joint arthroplasties started in 2013 in the LROI (Dutch Arthroplasty Register). A specific set of PROMs for THA and TKA patients are recommended nationally by the NOV to be

administered at several time points (pre-surgery and post-surgery). Currently, a standard outcome set mandated by the Dutch Orthopedic Society (NOV) comprises 20 items distributed across four different PROMs. As a result, an enormous amount of data is collected by each orthopedic healthcare institution. However, the data collection leads to the use of resources, costs, and difficulties associated with the data collection itself. This leads to an increased burden placed on patients and healthcare providers, as evidenced by patient survey fatigue manifesting by reduced completion rates and declining engagement over time[29].

Recommendation #6: reduce burden with short PROMs / PROMs using CAT

An advantage of PROMIS, similar to our findings in Chapters 4 and 5, is the reduction of the number of items questioned. This leads to a reduced burden on patients and healthcare suppliers[9]. This reduction can result in decreased data collection and a smaller carbon footprint.

III. Outer setting – the stakeholders

Facilitator/Barrier: stakeholder incentives and priorities

An important facilitator for the implementation of PROMs in the Netherlands is the national program *Uitkomstgerichte Zorg* ("Outcome-Based Care"). This program provides a strategic framework and shared vision for the systematic use of PROMs across healthcare settings. Within this context, a standardized set of PROMs (the Generic PROM set) has been developed and is nationally recommended for use in medical specialist care. This policy-driven standardization serves as a strong facilitator by providing clear guidance for healthcare institutions and ensuring comparability of outcomes across organizations. Furthermore, the Dutch National Health Care Institute (Zorginstituut Nederland) includes PROM as a mandatory quality indicator that help monitor and improve healthcare performance. The NOV provides data through the LROI, with an incentive to maintain a valid and comprehensive register of all arthroplasties, including PROM data. These three stakeholders have to align the decision-making regarding PROMs with clinical practice guidelines, rather than permitting external agencies to impose PROMs on organizations. The 'Wet kwaliteitsregistraties zorg' (2025) is expected to accelerate data collection by making the submission of (PROM) data mandatory by care providers to the quality registries.

Barrier: data collection

The Dutch Arthroplasty Register (LROI) has to collect and standardize data from all hospitals and medical clinics in the register. This process creates a significant administrative burden and requires substantial coordination to ensure consistency and accuracy, resulting in institutional resource demands. Furthermore, long-term follow-up can deliver unnecessary data[30]. Second, the data collection, transmission and storage creates a substantial carbon footprint.

Recommendation # 7: incentive towards outcome-based improvement

Achieving a return on investment through improved quality of PROMs and cost savings from reduced clinician time or VBHC initiatives will drive outcome-based care improvement. Overemphasis on PROM completion could inadvertently shift focus to narrowly target these metrics; therefore, incentive structures should be embedded that reward continuous quality improvement, rather than solely data reporting. Beyond the mandatory feedback to Zorginstituut Nederland, a more meaningful incentive in collecting PROMs is to assess healthcare institutions and professionals based on the implementation of demonstrable improvement initiatives grounded in established guidelines, outcome information, and efficiency. Such an approach would promote learning, enhance service delivery and provide proven effective care that benefits all patients. This reflects a clear trend towards transitioning the Dutch system from volume-based to outcome-based reimbursement. For example, the Integraal Zorgakkoord (IZA) aims to integrate outcome data into reimbursement models. To this end, a reassessment of NOV, LROI and Zorginstituut Nederland towards alternative strategies is warranted-especially given the funding from VWS allocated for outcome-oriented care initiatives. Facilitating dialogue among NOV, LROI, and insurers could lead to the creation of transparent incentive models—for example, bonus payments for high PROM completion rates or value-based reimbursement linked to patient outcomes. Another option used in the USA, is implementing an ICD-10 code for an electronic PROM consult.

IV. Characteristics of individuals – the roles and characteristics of individuals

Barrier: clinicians view

Challenges in PROM implementation derive also from clinicians' varied perceptions regarding their value. Many clinicians perceive PROMs as an administrative burden rather than recognizing them as beneficial tools for patient care[31]. Other key barrier

themes for orthopedic surgeons are logistical issues and difficulty interpreting and understanding PROM data[32].

Recommendation #8: leadership engagement and education

Engaging early adaptor clinicians as PROM ambassadors may transform perceptions among other clinicians, shifting their view of PROMs from administrative task to recognizing them as meaningful clinical data[33]. Furthermore, it is essential that surgeons receive education and that data is presented in an easily accessible and user-friendly manner within medical records[32,34].

Barrier: patient response rate

Even with valid PROMs measuring physical functioning, low response rates can render results unrepresentative, unreliable, or not useful. In the Netherlands, the response rate for THA and TKA in 2023 were as low as 65 and 56% pre-surgery, and 32 and 27% respectively at 12 months post-surgery (LROI, 2024a, 2024b). Acceptable response rate thresholds are often arbitrary and vary between 60 to 80%[35]. The low response rates causes selection bias, and therefore doubt about the representativeness of the data[36,37]. Response rates are influenced by patient demographics and create non-responder bias. Every effort should be made to increase capture rate and concise routine patient outcome measurement to avoid selection bias.

Recommendation # 9: patient engagement

Enhancing patient engagement is essential for improving response rates. Patients must clearly understand how completing PROMs can enhance their care and inform discussions with their primary care providers[38]. Other strategies for optimizing response rates include minimizing the number of items included in routine outcome measurement and ensuring those items are meaningful to patients[35,38]. Individual PROM questions which are relevant to the THA and TKA populations are likely to improve response rates and less burden leads to higher response rates. Furthermore, utilizing PROMIS allows patients with multiple chronic conditions to avoid completing separate PROMs for each specialty. If the current routine patient outcome measurement, with a regular follow-up to 12 months post-surgery, leads to a lower response rate, it is important to consider alternatives. For instance, serial measurement and targeted follow-up for non-responders for research purposes to improve response rates. However, when PROMs are used for quality improvement, all THA and TKA patients should be evaluated[35]. Using for example the KLIK PROM portal (www.hetklikt.nu) and EPIC (electronic health record), provides several educational activities for patients such

as flyers, information letters and videos[39] aimed at optimizing response rates through patient education.

V. Process – the activities and strategies used to implement the PROMs

Facilitator: broad stakeholder engagement

Broad stakeholder engagement facilitates the implementation of PROMs[33]. Significant efforts have already been made concerning Dutch PROM implementation strategies, and existing reports can serve as valuable guides, such as those from Zorginstituut Nederland, the Santeon hospitals and the NOV[40–42]. A comprehensive international guideline for implementing PROMs in practice is the PROTEUS Guide to Implementing PROs in Clinical Practice, which helps to design, implement and manage PRO systems in clinical care[43,44]. These reports should be used as clear protocol for structured implementation.

Recommendation #10: training support and outcome evaluation

After implementation of the new standard outcome set, the NOV should offer targeted training modules demonstrating how PROMs can facilitate shared decision-making and comparative benchmarking, thereby facilitating both clinicians and patients engagement. The LROI dashboards will continue to facilitate public sharing of hospital-level response rates and outcomes, encouraging research, transparency and motivation. Evaluation of the newly established core outcome set is essential to assess its effectiveness and utility.

Conclusion

This thesis is not merely a research endeavor but a call for practice improvement. Although its primary focus is methodological, the implications of validated PROMs extend to enhancing shared decision-making, tracking treatment effectiveness, and promoting value-based healthcare. Efforts should be focused on composing a standard set of PROMs that accurately reflect patients' physical function, quality of life and pain. It is up to the stakeholders to weigh the benefits of PROMIS versus the legacy measures. Let us strive for making patient-reported outcomes reliable, valid, accessible, and not burdensome, while ensuring their clinical implementation to demonstrate value. The transition to PROMIS could be part of this, since we consider PROMIS to be the preferred instrument for measuring physical function in THA and TKA patients due to its universal applicability, strong psychometric properties and better feasibility. The implications of

valid PROMs extend beyond measurement. Valid PROMs integrated within VBHC facilitate the identification and reduction of low-value or unnecessary interventions and PROM-based remote monitoring and telehealth approaches further reduce the environmental impact of care delivery. This will contribute to the reduction of the biggest problems in healthcare, the healthcare shortage and the carbon footprint. This thesis is one step in a broader initiative to achieve and implement a standard set of PROMs that measure adequate what truly matters to THA and TKA patients.

References

1. Damschroder LJ, Reardon CM, Widerquist MAO, Lowery J. The updated Consolidated Framework for Implementation Research based on user feedback. Implement Sci. 2022;17(1):75.
2. Braaksma C, Wolterbeek N, Veen MR, Prinsen CAC, Ostelo RWJG. Systematic review and meta-analysis of measurement properties of the Hip disability and Osteoarthritis Outcome Score - Physical Function Shortform (HOOS-PS) and the Knee Injury and Osteoarthritis Outcome Score - Physical Function Shortform (KOOS-PS). Vol. 28, Osteoarthritis and Cartilage. 2020;28(12):1525–38.
3. Braaksma C, Wolterbeek N, Veen RMR, Prinsen CAC, Ostelo RWJG. The Hip Disability and Osteoarthritis Outcome Score-Physical Function Shortform Does Not Adequately Represent Physical Functioning in Patients Undergoing Total Hip Arthroplasty. Value Heal. 2022;25(11):1894-901.
4. Werkgroep WMNProm. NOV PROMs-advies orthopedie 2020. 2020.
5. Rolfson O, Wissig S, van Maasakkers L, Stowell C, Ackerman I, Ayers D, et al. Defining an International Standard Set of Outcome Measures for Patients With Hip or Knee Osteoarthritis: Consensus of the International Consortium for Health Outcomes Measurement Hip and Knee Osteoarthritis Working Group. Arthritis Care Res [Internet]. 2016 Nov 1 [cited 2019;68(11):1631–9. Available from: http://doi.wiley.com/10.1002/acr.22868
6. Gagnier JJ, Huang H, Mullins M, Marinac-Dabić D, Ghambaryan A, Eloff B, et al. Measurement properties of patient-reported outcome measures used in patients undergoing total hip arthroplasty: A systematic review. JBJS Rev. 2018;6(1).
7. Holmenlund C, Overgaard S, Bilberg R, Varnum C. Evaluation of the Oxford Hip Score: Does it still have content validity? Interviews of total hip arthroplasty patients. Health Qual Life Outcomes. 2021;19(1):237.
8. Gagnier JJ, Mullins M, Huang H, Marinac-Dabic D, Ghambaryan A, Eloff B, et al. A Systematic Review of Measurement Properties of Patient-Reported Outcome Measures Used in Patients Undergoing Total Knee Arthroplasty. Journal of Arthroplasty. 2017;32(5):1688-97.e7..
9. Braaksma C, Wolterbeek N, Veen MR, Poolman RW, Pronk Y, Klaassen AD, et al. Assessing the measurement properties of PROMIS Computer Adaptive Tests, short forms and legacy patient reported outcome measures in patients undergoing total hip arthroplasty. J Patient-Reported Outcomes. 2024;8(1):121.
10. Thomas ML. The Value of Item Response Theory in Clinical Assessment: A Review. Assessment. 2011;18(3):291–307.
11. Brodke DJ, Hung M, Bozic KJ. Item Response Theory and Computerized Adaptive Testing for Orthopaedic Outcomes Measures. J Am Acad Orthop Surg. 2016;24(11):750–4.
12. Cappelleri JC, Jason Lundy J, Hays RD. Overview of Classical Test Theory and Item Response Theory for the Quantitative Assessment of Items in Developing Patient-Reported Outcomes Measures. Clin Ther. 2014;36(5):648–62.
13. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. The patient-reported outcomes measurement information system (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. J Clin Epidemiol. 2010;63(11):1179–94.
14. Terwee C, Ahmed S, Alhasani R, Alonso J, Bartlett S, Chaplin J, et al. Comparable real-world patient-reported outcomes data across health conditions, settings, and countries: the PROMIS international collaboration. NEJM Catal. 2024;5(9).
15. Smith TO, Hawker GA, Hunter DJ, March LM, Boers M, Shea BJ, et al. The OMERACT-OARSI Core Domain Set for Measurement in Clinical Trials of Hip and/or Knee Osteoarthritis. J Rheumatol. 2019;46(8):981–9.
16. Abma IL, Butje BJD, ten Klooster PM, van der Wees PJ. Measurement properties of the Dutch–Flemish patient-reported outcomes measurement information system (PROMIS) physical function item bank and instruments: a systematic review. Health Qual Life Outcomes. 2021;19(1):62.
17. Oude Voshaar MAH, ten Klooster PM, Glas CAW, Vonkeman HE, Taal E, Krishnan E, et al. Validity and measurement precision of the PROMIS physical function item bank and a content validity–driven 20-item short form in rheumatoid arthritis compared with traditional measures. Rheumatology. 2015;54(12):2221-9.

18. Czerwonka N, Gupta P, Desai SS, Hickernell TR, Neuwirth AL, Trofa DP. Patient-reported outcomes measurement information system instruments in knee arthroplasty patients: a systematic review of the literature. Knee Surg Relat Res. 2023;35(1):27.
19. Kagan R, Anderson MB, Christensen JC, Peters CL, Gililland JM, Pelt CE. The Recovery Curve for the Patient-Reported Outcomes Measurement Information System Patient-Reported Physical Function and Pain Interference Computerized Adaptive Tests After Primary Total Knee Arthroplasty. J Arthroplasty. 2018;33(8):2471–4.
20. Zonjee VJ, Abma IL, de Mooij MJ, van Schaik SM, Van den Berg-Vos RM, Roorda LD, et al. The patient-reported outcomes measurement information systems (PROMIS®) physical function and its derivative measures in adults: a systematic review of content validity. Qual Life Res. 2022;31(12):3317–30.
21. Segawa E, Schalet B, Cella D. A comparison of computer adaptive tests (CATs) and short forms in terms of accuracy and number of items administrated using PROMIS profile. Qual Life Res. 2020;29(1).
22. Werkgroep Generieke PROMs [Internet]. Available from: https://demedischspecialist.nl/sites/default/files/2022-02/adviesrapport_werkgroep_generieke_proms.pdf
23. Oude Voshaar M, Terwee CB, Haverman L, van der Kolk B, Harkes M, van Woerden CS, et al. Development of a standard set of PROs and generic PROMs for Dutch medical specialist care. Qual Life Res. 2023;32(6):1595–605.
24. Braaksma C, Mobers BGN, van Beers LWAH, Pronk Y, Klaassen AD, Poolman RW, et al. Validating existing crosswalks between legacy PROMs and PROMIS measuring physical functioning in patients undergoing total hip and total knee arthroplasty. Adv Patient-Reported Outcomes. 2025;100201.
25. Stephan A, Stadelmann VA, Leunig M, Impellizzeri FM. Measurement properties of PROMIS short forms for pain and function in total hip arthroplasty patients. J Patient-Reported Outcomes. 2021;5(1):41.
26. Prinsen CAC, Vohra S, Rose MR, Boers M, Tugwell P, Clarke M, et al. How to select outcome measurement instruments for outcomes included in a "Core Outcome Set" - a practical guideline. Trials. 2016;17(1):449
27. 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;63(7):737–45.
28. Glenwright BG, Simmich J, Cottrell M, O'Leary SP, Sullivan C, Pole JD, et al. Facilitators and barriers to implementing electronic patient-reported outcome and experience measures in a health care setting: a systematic review. J Patient-Reported Outcomes. 2023;7(1):13.
29. Papuga MO, Dasilva C, McIntyre A, Mitten D, Kates S, Baumhauer JF. Large-scale clinical implementation of PROMIS computer adaptive testing with direct incorporation into the electronic medical record. Heal Syst. 2018;7(1).
30. Spece H, Kurtz MA, Piuzzi NS, Kurts SM. Patient-reported outcome measures offer little additional value two years after arthroplasty a systematic review and meta-analysis. Bone Joint J. 2025;107B(3).
31. Shapiro LM, Spindler K, Cunningham B, Koh J. Patient-Reported Outcome Measure Collection and Utilization: A Survey of American Academy of Orthopaedic Surgeons Members. J Am Acad Orthop Surg. 2024;32(3):114–22.
32. Heath EL, Harris IA, Romero L, Ackerman IN. A systematic review of qualitative studies examining barriers and facilitators to orthopaedic surgeon engagement with patient-reported outcome measures data. J Patient-Reported Outcomes. 2024;8(1):144.
33. Anderson M, van Kessel R, Wood E, Stokes A, Fistein J, Porter I, et al. Understanding factors impacting patient-reported outcome measures integration in routine clinical practice: an umbrella review. Qual Life Res. 2024;33(10):2611–29.
34. Foster A, Croot L, Brazier J, Harris J, O'Cathain A. The facilitators and barriers to implementing patient reported outcome measures in organisations delivering health related services: a systematic review of reviews. J Patient-Reported Outcomes. 2018;2(1):46.
35. Rolfson O., Eresian Chenok K., Bohm E., Lubbeke A., Denissen G., Dunn J., Lyman S., Franklin P., Dunbar M., Overgaard S., Garellick G., Dawson J.. Patient-reported outcome measures in arthroplasty registries: Report of the Patient-Reported Outcome Measures Working Group of the International Society of Arthroplasty Registries: Part I. Overview and rationale for patient-reported outcome measures. Acta Orthop [Internet]. 2016;87:3–8. Available from: http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L610356344

36. Geilen JEJ., Hoelen TA, Schotanus MGM, Spekenbrink-Spooren A, Boonen B, Most J. Defining Clinically Meaningful Thresholds for 12-Month Patient-Reported Outcomes in Total Hip Arthroplasty; Toward Improving Threshold Accuracy. Arthroplast Today. 2025;32:101649.
37. Stephens AR, Bender NR, El-Hassan R, Patel RK. Evidence of non-response bias in patient reported outcome measurement information system surveys. Interv Pain Med. 2025;4(2):100588.
38. Unni E, van Muilekom MM, Absolom K, Bajgain B, Haverman L, Santana M. Educating patients about patient-reported outcomes—are we there yet? J Patient-Reported Outcomes. 2024;8(1):113.
39. Haverman L, van Oers HA, Limperg PF, Hijmans CT, Schepers SA, Sint Nicolaas SM, et al. Implementation of Electronic Patient Reported Outcomes in Pediatric Daily Clinical Practice: The KLIK Experience. Clin Pract Pediatr Psychol. 2014;2(1):50–67.
40. Baalen M van, Gommans T, Berens M. PROMs in de spreekkamer- Succes- en faalfactoren en lessen voor implementatie. 26-10-2018.
PROMs+in+de+spreekkamer+Succesfactoren+en+faalfactoren+en+lessen+voor+implementatie.pdf
41. Proms-advies NOV PROMs-advies orthopedie 2020.
42. Santeon. stappenplan PROMs implementatie.
43. Crossnohere NL, Anderson N, Baumhauer J, Calvert M, Esparza R, Gulbransen S, et al. A framework for implementing patient-reported outcomes in clinical care: the PROTEUS-practice guide. Nat Med. 2024;30(6):1519–20.
44. The PROTEUS Guide to Implementing Patient-Reported Outcomes in Clinical Practice: A Synthesis of Resources (the PROTEUS-Practice Guide).

CHAPTER 8

Summary

Introduction

Chapter 1 provides a general introduction to this thesis. Patient-Reported Outcome Measures (PROMs) are designed to capture patients' perspectives and are mandated quality indicators in the Netherlands for all orthopedic clinics performing total hip arthroplasty (THA) and total knee arthroplasty (TKA). The Dutch Orthopedic Society (NOV) recommends a specific set of PROMs for patients undergoing THA and TKA. However, these so-called legacy PROMs have substantial limitations in their measurement properties. Furthermore, PROMs are often poorly implemented in orthopedic care.

Optimizing PROM selection and use is essential for improving patient-centered orthopedic care, research, and policy-making. Insufficient implementation or inadequate PROMs may lead to outcomes that do not adequately reflect what matters to patients and inefficient and less patient-centered healthcare. Therefore, the overarching aim of this thesis is to optimize the measurement of physical function and pain in THA and TKA patients using PROMs.

This thesis is structured around three themes:

Part I. Measurement properties of legacy PROMs evaluating physical function in THA and TKA.

Part II: Towards an adequate alternative patient-reported outcome measure in THA and TKA.

Part III: Standardizing legacy PROM score conversions towards PROMIS scores.

Part I. Measurement properties of legacy PROMs evaluating physical function in THA and TKA

Part 1 comprises two chapters evaluating the psychometric properties of legacy PROMs measuring physical function in THA and TKA. Chapter 2 presents a systematic review of the measurement properties of the Hip disability and Osteoarthritis Outcome Score - Physical function Shortform (HOOS-PS) and the Knee Injury and Osteoarthritis Outcome Score - Physical function Shortform (KOOS-PS). These PROMs were, at the time of the

study, recommended by the NOV for evaluating THA and TKA. Chapter 3 examined the content validity of the HOOS-PS by interviewing patients and clinicians.

Chapter 2

Chapter 2 presents a systematic review of the measurement properties of the HOOS-PS and KOOS-PS in patients undergoing THA and TKA. These legacy PROMs were selected as outcome measurement instruments by global standard sets of outcome measures, arthroplasty registries and clinical research studies. The study was conducted according to the COSMIN guideline for systematic reviews of PROMs. The most notable finding in this review of 23 articles was the inconsistent and often insufficient evidence for content validity, suggesting that scores on the HOOS-PS and KOOS-PS may insufficiently reflect physical functioning. Furthermore, we found evidence for insufficient construct validity and responsiveness in patients with knee osteoarthritis receiving conservative treatment. Consequently, the chapter concludes that the use of the HOOS-PS and KOOS-PS for outcome comparison, evaluation of treatment effects, or benchmarking in patients with hip or knee complaints or undergoing arthroplasty should only be done with great caution.

Chapter 3

In Chapter 3, a combined quantitative and qualitative research approach was used to assess the content validity of the HOOS-PS. The HOOS-PS is a frequently used PROM for assessing physical functioning in patients with hip problems. This study aimed to assess the content validity of the HOOS-PS. 51 patients and 25 experts completed questionnaires regarding the relevance, comprehensiveness and comprehensibility of the HOOS-PS and 5 semi-structured interviews explored issues in depth identified in the quantitative data. Thematic content analysis was conducted using a coding frame. Only one of the five items was considered relevant for measuring physical functioning according to patients and experts. Comprehensiveness and comprehensibility were considered insufficient. Several items were found to be ambiguous or double-barreled. This study raised concerns about the content validity of the HOOS-PS: the majority of the items were considered not relevant, the HOOS-PS was considered not comprehensive, and several items were considered not comprehensible. These findings challenge the applicability of the HOOS-PS in clinical practice, research, VBHC, and benchmarking.

Part II: Towards an adequate alternative patient-reported outcome measure in THA and TKA

Given the limitations identified in Part 1, Part 2 explores an innovative alternative: the Patient-Reported Outcomes Measurement Information System (PROMIS®). This part consists of two chapters comparing the psychometric properties with those of legacy PROMs that evaluate physical function and pain. Chapter 4 assessed and compared the measurement properties in THA patients, and Chapter 5 assessed them in TKA patients.

Chapter 4

Legacy PROMs evaluating outcomes of total hip arthroplasty (THA) have several limitations regarding their measurement properties and interpretation of scores. One innovation in PROM development is the use of Computerized Adaptive Testing (CAT). PROMIS is a validated system based on Item Response Theory, and enables the use of CAT. In Chapter 4, the measurement properties of PROMIS and legacy instruments in patients undergoing THA were assessed. In this multicenter study, 208 patients completed a questionnaire twice, including Dutch-Flemish PROMIS v1.2 Physical Function (PROMIS-PF), v1.1 Pain Interference (PROMIS-PI) CATs and short forms, PROMIS v1.0 Pain Intensity, and legacy PROMs (Hip disability and Osteoarthritis Outcome Score (HOOS), HOOS-Physical function Shortform (HOOS-PS), Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), Oxford Hip Score (OHS), and two numeric rating scales measuring pain). The reliability, measurement precision (Standard Error of Measurement (SEM)), smallest detectable change (SDC), and burden of PROMIS instruments were presented head-to-head to legacy PROMs. Furthermore, construct validity was assessed. The PROMIS-PF was found to be less burdensome, with high measurement precision, almost no minimal or maximal scores, and equal reliability compared to legacy instruments measuring physical functioning in patients undergoing THA. The PROMIS Pain Intensity 1a was found to be comparable to the legacy pain instruments in terms of burden, reliability, and SDC. Measuring the construct Pain Interference may not have additional value in this population because of its high correlation with instruments measuring physical functioning. The SDC values presented in this study can be used for individual patient monitoring.

Chapter 5

Chapter 5 assessed the psychometric properties of PROMIS measures compared to legacy PROMs. In this multicenter test-retest study, 210 patients were included from three orthopedic departments. Patients completed a questionnaire including PROMIS CAT, PROMIS SF, and legacy PROMs measuring physical function and pain. Measurement precision (SEM), smallest detectable change (SDC), construct validity, burden and

extreme scores were investigated. All PROMIS CAT, SFs, and legacy PROMs showed adequate test-retest reliability (ICCs between .74 and .94). The measurement range of PROMIS measures was 1-81 T-score points, and of legacy PROMs 0-100. Regarding physical function, the SEM varied between 1.6-2.1 for PROMIS and 3.6-6.9 for legacy PROMs. The SDC varied between 2.8-6.1 for PROMIS and 7.3-19.9 for legacy PROMs. PROMIS showed no extreme scores using 5-20 items, legacy PROMs showed 1.2 – 3% extreme scores using 7-17 items. Regarding Pain (Interference), the SEM varied between 1-2.1 for PROMIS and 8.1-12.8 for legacy PROMs. The SDC varied between 2.8-6.1 for PROMIS and 22.5-35.4 for legacy PROMs. PROMIS showed 0-9.5% extreme scores using 5-8 items, legacy PROMs 8.7-19.1% extreme scores using 1-9 items. The construct validity was sufficient for all PROMIS CAT and SF. Hence, the results of this study showed that PROMIS measures seem more efficient for assessing patient-reported physical function in TKA, offering reduced burden and measurement error, and minimizing the occurrence of extreme scores. This can facilitate more accurate and patient-centered evaluations. This study may support a potential shift from legacy PROMs toward PROMIS in patients undergoing TKA.

Part III: Standardizing legacy PROM score conversions towards PROMIS scores

Part III evaluates existing crosswalks for converting legacy PROM scores in THA and TKA patients to PROMIS instruments, supporting data continuity during a transition to PROMIS.

Chapter 6

Chapter 6 validated existing crosswalks for transforming scores of the HOOS-PS, the KOOS-PS, and the KOOS-ADL subscale to the PROMIS Physical Function metric. 422 patients from three orthopedic departments completed online questionnaires including the aforementioned PROMs. After converting the legacy PROM scores to the PROMIS metric using existing crosswalk tables, Pearson's correlation and the intraclass correlation coefficient (ICC) between predicted (based on the crosswalk) and observed PROMIS scores were calculated. The level of agreement between the predicted and observed PROMIS scores was assessed using a Bland-Altman plot, and the Limits of Agreement (LoA) were calculated. Last, the percentage of patients for whom the predicted score was considered acceptably comparable to the observed score (difference <2 points) was determined. The study results showed adequate correlations (≥0.70) and ICC's (≥0.70) between observed and predicted PROMIS scores, indicating

good performance of the crosswalks and suggesting good agreement at the group level. Mean differences between predicted and observed PROMIS T-scores based on the HOOS-PS, KOOS-PS, and KOOS ADL were respectively 1.4, -0.3, and 1.0. The LoA varied between -10.8 and 12.4 T-score points, indicating substantial differences between the observed and predicted PROMIS scores at the individual patient level. Only 25.7%-39.8% of the patients had a predicted PROMIS score that was acceptably comparable to the observed PROMIS score. This study concluded that the existing HOOS-PS, KOOS-PS, and KOOS-ADL crosswalks towards PROMIS PF seem to be appropriate for group-level use but are not suitable for individual-level predictions of PROMIS scores.

Discussion

In Chapter 7 (Discussion), key facilitators and barriers for PROM implementation in orthopedic care were identified using the Consolidated Framework for Implementation Research (CFIR[1]). Furthermore, recommendations for solutions were presented. The results highlight the need for composing a standard set of PROMs that accurately reflect patients' physical function and pain. The transition to PROMIS could be part of the solution, due to its broad applicability, strong psychometric properties and better feasibility. Furthermore, effective integration of PROMs into clinical practice and Value Based Healthcare (VBHC) frameworks, can facilitate the identification and reduction of low-value or unnecessary interventions and PROM-based remote monitoring and telehealth approaches. Patient, leader- and stakeholder engagement may optimize response rates and implementation. This potentially contributes to the reduction of healthcare resource use and carbon footprint. Implementing a standard set of PROMs that measure adequate what truly matters to THA and TKA patients can enhance shared decision-making, monitoring of treatment effectiveness, and more sustainable orthopedic healthcare.

CHAPTER 9

Curriculum Vitae
PhD Portfolio
List of publications
List of author contributions
Dankwoord

Curriculum vitae

Christel Braaksma was born on August 31, 1990 in Ens, the Netherlands, where she grew up with her parents and sister, and later welcomed also her stepfather, stepbrother, and stepsister. She completed the Athenaeum at Zuyderzee College in 2007. In 2011, she obtained a bachelor's degree in Human Movement Sciences from the University of Groningen. During her final year, she enrolled in the pre-master's program for medical school at the same university. Following completion of her bachelor's degree, she undertook a clinical rotation year in Deventer and Utrecht and graduated from medical school in 2015.

After graduation, Christel began her professional journey as a physician at St. Antonius Hospital, initially working in the department of Orthopedic Surgery before transitioning to the department of General Surgery. In 2018, Christel commenced her residency training in orthopedic surgery. Her training was carried out at the department of Surgery (head: Dr. Boerma), the orthopedic departments of St. Antonius Hospital (head: Dr. Veen), UMC Utrecht (head: Dr. Van der Wal), OLVG (head: Dr. Van Deurzen), and the Trauma department of UMC Utrecht (head: Dr. Houwert). She is expected to complete her orthopedic specialty training at St. Antonius Hospital at the end of 2026, under the supervision of Dr. Van Dijk.

During her residency, Christel initiated and coordinated a prospective multicenter study examining the measurement properties of Patient-Reported Outcome Measures in patients undergoing total hip and total knee arthroplasty, for which she obtained research funding from the Dutch Arthroplasty Register (LROI) and the St. Antonius Research fund. She has contributed to several peer-reviewed articles and has presented her research at multiple national and international conferences.

Christel currently lives in Utrecht with her partner Dennis van der Linden, and their children Veda (2024) and Huub (2026).

PhD Portfolio

PhD Portfolio

Affiliation: Vrije Universiteit Amsterdam
Faculty: Faculteit der Bèta wetenschappen
Graduate school: GS bèta – Gezondheidswetenschappen
Institute: Amsterdam Movement Sciences
PhD period: June 2021 – Dec 2025

PhD supervisors: prof. dr. R.W.J.G. Ostelo
prof. dr. C.B. Terwee

PhD co-supervisors: dr. N. Wolterbeek
dr. M.R. Veen

Education	Location	ECTs
Courses		
Research Integrity	Online Epigeum, 2024	2
Writing a scientific article	VU Taalcentrum, 2018	3
Epidemiologisch onderzoek: basisprincipes	Amsterdam UMC, EpidM	4
Clinimetrics	VUmc	5
WMO Good Clinical Practice	Online My GCP, 2024	2
MSH meetings	AMC, Amsterdam Movement Sciences	1
Seminars		
Working group VBHC Osteoarthritis of the hip	Antonius Hospital, 2021-2024	1
Working group Santeon Zorg Bij Jou	Santeon	2
Conferences		
Oral Promis Health Organization	Prague	2
2 poster presentations PROMIS Health Organization	Prague	2
Attending conferences and symposia	Varying	2,14
Organization scientific symposium Food for Thought	Nieuwegein	1
Other		
Grant application and funding LROI €78.599	2021	2
Teaching activities		ECTs
Lectures		
Lectures bachelor degree medical students, nurses and OR nurses	UMC Utrecht, OLVG, VUmc, Antonius Hospital	2
Supervision		
Head medical intern supervisor, supervising research students	UMC Utrecht, Antonius Hospital	1
Total activities		**32,14 ECs**

List of publications

List of publications

This thesis

- C. Braaksma, N. Wolterbeek, M.R. Veen, C.A.C. Prinsen, R.W.J.G. Ostelo. Systematic review and meta-analysis of measurement properties of the Hip disability and Osteoarthritis Outcome Score - Physical function Shortform (HOOS-PS) and the Knee Injury and Osteoarthritis Outcome Score - Physical Function Shortform (KOOS-PS). *Osteoarthritis and cartilage*. 2020 Dec; 28(12): 1525-1538.
- C. Braaksma, N. Wolterbeek, M.R. Veen, C.A.C. Prinsen, R.W.J.G. Ostelo. The Hip Disability and Osteoarthritis Outcome Score-Physical Function Shortform Does Not Adequately Represent Physical Functioning in Patients Undergoing Total Hip Arthroplasty. *Value in Health.* 2022 Nov; 25(11): 1894-1901.
- C. Braaksma, N. Wolterbeek, M.R. Veen, R.W. Poolman, Y. Pronk, A.D. Klaassen, R.W.J.G. Ostelo, C.B. Terwee. Assessing the measurement properties of PROMIS Computer Adaptive Tests, short forms and legacy patient reported outcome measures in patients undergoing total hip arthroplasty. *Journal of Patient-Reported Outcomes*. 2024 Oct; 8:121.
- C. Braaksma, N. Wolterbeek, M.R. Veen, R.W. Poolman, Y. Pronk, A.J. Rasker, R.W.J.G. Ostelo, C.B. Terwee. A comparison of the psychometric properties of PROMIS computer adaptive tests and short forms versus legacy patient-reported outcome measures in total knee arthroplasty patients. Arthroplasty Today 2026 38; in press.
- C. Braaksma, B.G.N. Mobers, L.W.A.H. van Beers, Y. Pronk, A.D. Klaassen, R.W. Poolman, B.D. Schalet, D. Delawi, R.W.J.G. Ostelo, C.B. Terwee. Validating existing crosswalks between legacy PROMs and PROMIS measuring physical functioning in patients undergoing total hip and total knee arthroplasty. *Advances in Patient-Reported Outcomes.* 2025 Oct; in press

Other publications

- C. Braaksma, N. Wolterbeek, M.R. Veen. Survival, complications and outcomes of the Birmingham Hip Resurfacing compared to cementless total hip arthroplasty. *International Journal of Orthopaedics*. 2018; 5(2): 1-5
- C. Braaksma, D. Vermeulen, A.M. van Leeuwen, M.J.G.M. Speth, T.M. Piscaer. Bilateral vanishing hips, coincidence or systemic disease? A case report and overview of current literature. *Journal of Orthopaedics*. 2018;15(2): 641-644

- C. Braaksma, V. Oehlers, M.R. Veen, N. Wolterbeek. Patient characteristics do not predict the change in physical functioning following arthroplasty measured by the HOOS-PS and KOOS-PS. *Journal of Orthopaedics*. 2020 Jan 10;20:122-124.
- C. Braaksma, D. Boerma. Een buitengewoon seksistisch stukje tekst. *NTVH*. Maart 2020; 29(3): 24.
- C. Braaksma, J. Otte, R.N. Wessel, N. Wolterbeek. Investigation of the efficacy and safety of ultrasound-standardized autologous blood injection as treatment for lateral epicondylitis. *Clinics in Shoulder and Elbow*. 2022 Mar; 25(1):57-64.

List of author contributions

List of author contributions

1. Systematic review and meta-analysis of measurement properties of the Hip disability and Osteoarthritis Outcome Score - Physical Function Shortform (HOOS-PS) and the Knee Injury and Osteoarthritis Outcome Score - Physical Function Shortform (KOOS-PS) *Osteoarthritis and Cartilage* 2020 Dec; 28(12):1525-1538

   Authors: C. Braaksma, N. Wolterbeek, M.R. Veen, C.A.C. Prinsen, R.W.J.G. Ostelo
   Author contributions: Conception and design of the study: CB, NW, SP, RV, RO; Analysis and interpretation of the data: CB, NW, SP, RV, RO; Drafting of the article: CB, NW; Critical revision of the article for important intellectual content: CB, NW, SP, RV, RO, Dr. C.B. Terwee; Final approval of the article: CB, NW, SP, RV, RO; statistical expertise: CB, NW, SP, RO, Prof. Dr. HCW de Vet; Collection and assembly of data: CB, NW, SP.

2. The Hip Disability and Osteoarthritis Outcome Score-Physical Function Shortform Does Not Adequately Represent Physical Functioning in Patients Undergoing Total Hip Arthroplasty *Value in Health* 2022 Nov; 25(11):1894-1901

   Authors: C. Braaksma, N. Wolterbeek, M.R. Veen, C.A.C. Prinsen, R.W.J.G. Ostelo
   Value in Health *2022; 25(11):1894-1901*
   Author contributions: Concept and design of the study: CB, NW, RV, SP, RO; Acquisition of data: CB, NW, RV; Analysis and interpretation of data: CB, NW, RV, SP, RO. Drafting of the manuscript: CB, NW, RV. Critical revision of the paper for important intellectual content: NW, RV, SP, RO. Statistical analysis: CB, NW, RV, SP, RO. Provision of study materials or patients: CB, NW, RV. Administrative, technical, or logistic support: CB, NW, SP, RO. Supervision: NW, RV, SP, RO.

3. Assessing the measurement properties of PROMIS Computer Adaptive Tests, short forms and legacy patient reported outcome measures in patients undergoing total hip arthroplasty *Journal of Patient-Reported Outcomes* 2024 Oct 21;8(1):121

   Authors: C. Braaksma, N. Wolterbeek, M.R. Veen, R.W. Poolman, Y. Pronk, A.D. Klaassen, R.W.J.G. Ostelo, C.B. Terwee.

   Author contributions: C.B. and N.W. conceptualized, arranged the acquisition, analyzed and did the management and coordination of the project. C.B., N.W., Y.P. and A.D.K. collected data. C.B., N.W., M.R.V., R.W.P., Y.P., A.D.K., R.W.J.G.O. and

C.B.T. were involved in the design of the study. C.B. wrote the initial draft. N.W., M.R.V., R.W.P., Y.P., A.D.K., C.B.T. and R.W.J.G.O. reviewed, edited and supervised. All authors read and approved the final manuscript.

4. A comparison of the psychometric properties of PROMIS Computer Adaptive Tests and short forms versus legacy patient-reported outcome measures in total knee arthroplasty patients *Arthroplasty Today* 2026 38; in press.

   Authors: C. Braaksma, N. Wolterbeek, M.R. Veen, R.W. Poolman, Y. Pronk, A.J. Rasker, R.W.J.G. Ostelo, C.B. Terwee
   Author contributions: C.B. and N.W. conceptualized, arranged the acquisition, analyzed and did the management and coordination of the project. C.B., N.W., Y.P. and A.R. collected data. C.B., N.W., M.R.V., R.W.P., Y.P., A.R., R.W.J.G.O. and C.B.T. were involved in the design of the study. C.B. wrote the initial draft. N.W., M.R.V., R.W.P., Y.P., A.R., C.B.T. and R.W.J.G.O. reviewed, edited and supervised. All authors read and approved the final manuscript.

5. Validating existing crosswalks between PROMs and PROMIS measuring physical functioning in patients undergoing total hip and total knee arthroplasty *Advances in Patient-Reported Outcomes.* 2025 Oct 1; in press.

   Authors: C. Braaksma, B.G.N. Mobers, L.W.A.H. van Beers, Y. Pronk, A.D. Klaassen, R.W. Poolman, B.D. Schalet, D. Delawi, R.W.J.G. Ostelo, C.B. Terwee
   Author contributions: Y. Pronk: Writing – review & editing, Validation, Project administration, Methodology, Investigation, Data curation. A.D. Klaassen: Project administration, Methodology, Investigation, Data curation. R.W. Poolman: Writing – review & editing, Validation, Supervision. D. Delawi: Writing – review & editing, Supervision, Data curation. B.D. Schalet: Writing – review & editing, Validation, Supervision, Methodology, Formal analysis. R.W.J.G. Ostelo: Writing – review & editing, Validation, Supervision, Methodology, Conceptualization. C.B. Terwee: Writing – review & editing, Validation, Supervision, Methodology, Conceptualization. Braaksma Christel: Writing – original draft, Visualization, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. B.G.N. Mobers: Writing – original draft, Project administration, Formal analysis. L.W.A.H. van Beers: Writing – review & editing, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Conceptualization.

Dankwoord

Dankwoord

Dit hoofdstuk markeert het einde van mijn PhD. Wat ooit begonnen is als stage wetenschap in het Antonius, is tijdens mijn opleiding afgeschreven tot dit proefschrift. Hoewel mijn opleider vd Wal (UMCU) zegt; 'het enige echte promoveren is de promotie die verricht wordt in eigen tijd tijdens je opleiding', citeer ik liever mijn vader: 'het is maar goed dat je vooraf niet weet waar je aan begint'. Ik denk met dit proefschrift te hebben bijgedragen aan het meetbaar maken van pijn en fysiek functioneren van patiënten die een totale knie of heupprothese krijgen.

I am not so interested in how they move as in what moves them- Pina Bausch

In Nederland hebben we het geluk dat de integriteit van wetenschap in het algemeen nog wordt gerespecteerd. Ik verwacht dan ook dat mijn discussie bijdraagt aan de transitie naar het beter meten in deze patiëntenpopulatie. Dit proefschrift had ik nooit kunnen schrijven zonder mijn supervisors, collegae, vrienden en familie. Dank voor jullie bijdrage.

Speak the truth, write with clarity, and defend it to your very end - Ludwig Boltzmann

Geachte promotor prof. Ostelo, beste Raymond. Je bent zoals ik mij een professor had voorgesteld (behoudens de hiërarchie, daar houd je niet zo van): altijd zicht op het doel met behoud van de helicopterview. De immer opbeurende meetings, het relativeren en je prettige begeleiding zorgden ervoor dat ik het proefschrift af wilde maken. Ontzettend veel dank voor de gegeven energie en het vertrouwen, die hebben bijgedragen aan de totstandkoming van dit proefschrift.

Geachte promotor prof. Terwee, beste Caroline. Ik heb me vaak afgevraagd hoe iemand zoveel kan publiceren en zich toch nog met de interpunctie in de artikelen van de promovendus wil bemoeien. Je was een fantastische aanvulling; jouw methodologische kennis en je immer scherpe oog zijn een ware inspiratie geweest.

Geachte copromotor Dr. Veen, beste Remmelt, in het pas gebouwde Antonius Leidsche Rijn kwam ik bij jou solliciteren voor een wetenschappelijke stage. Het einde van dat verhaal ligt nu voor je. Je hebt altijd vertrouwen in me gehouden en me begeleid als een ware opleider. Bij 'ons' in het Anton voelde ik me altijd als een vis in het water. Ook heb je me naar het Anton gehaald toen ik even niet op mijn plek zat. Daar blijf ik je altijd

dankbaar voor. Overigens was dit boekje totaal niet nodig geweest als we jouw Cruijffiaanse waarheid volgen: '*Een nieuwe knie plaats je als de patient op is en de knie op is'.*

Geachte copromotor dr. Wolterbeek, lieve Nienke, in mijn oudste coschap kwam ik bij jou op de kamer werken. Aan jou de zware kluif mij *from scratch* alles bij te brengen... dank voor je engelengeduld. Door de jaren heen hebben we lief en leed gedeeld. Het was me een waar genoegen met je samen te werken -en jouw vrouwelijke *touch* te midden van het mannenbolwerk van de orthopedie was erg fijn.

Geachte leden van de beoordelingscommissie, geachte Prof. Dr. J.E. Bosmans, Dr. M.A.H. Oude Voshaar
Dr. D.O. Verbeek MBA, Prof. Dr. P.J. van der Wees, Prof. Dr. T.P.M. Vliet Vlieland, Prof. Dr. T. Gosens, hartelijk dank voor uw tijd en expertise bij het beoordelen van mijn proefschrift. Ik kijk er naar uit om met u van gedachten te wisselen over de inhoud.

Graag zou ik ook alle coauteurs bedanken die mee hebben gewerkt aan de verschillende publicaties in dit proefschrift, voor hun tijd en inspanningen.

Martijn, wat ben jij een warme opleider. Degene die op OK vraagt hoe het met je ouders gaat. Of je vader al geploegd heeft. En die met de MAKO liever egt dan ploegt (of was het andersom;)?). *Down-to-earth*, betrokken en met een hart voor de zaak en voor ons. Je begrijpt altijd wat het meest belangrijk is (niet werk). Dank voor alles.

Lieve Diyar, in alle jaren dat ik je nu alweer ken, ben je er altijd voor me - of het nu gaat om uitdagingen in de privésfeer of carrière struggles. Je daagt me altijd uit om slimmer en sterker te worden – *as strong as an ox and almost twice as clever*. Ik kijk uit naar het moment dat ik je in de ring te sterk af ben, want in het ziekenhuis ga ik het verliezen.

Fiona, jij bent een rots in de branding voor de maatschap, en een altijd warme vrouw met een luisterend oor te midden van veel alfamannen. Jouw betrokkenheid en enthousiasme bij mijn persoonlijke overwinningen (zowel een artikel als een baby) zijn mij heel dierbaar.

Lieve sandeep boys, aka het elite team: lieve Paul, Nick, Joost en Thom. Zonder de Strava stepsessies, corona party's onder (en in) het systeemplafond en de Liverpoolse brandwonden, was ik allang ingesukkeld tijdens mijn opleiding. Jullie geven de dagen

glans en zorgen ervoor dat ik weer zin heb in een dagje poli. Hoewel de vrijmibo's ons net zo goed afgaan. Nog veel te gaan hoop ik!

Orthopedisch chirurgen en ROGO-middenwest AIOS, dank voor de prachtige tijd in het OLVG, UMCU en Antonius. Ik heb genoten!

Pottertjes!!! Roos & Nadine(L) Met jullie hangend uit het raam met uitzicht op de Neude, Bieber-dansjes doen, date-leed delen. En inmiddels mama's die toch nog af en toe in de lampen willen hangen. Ik ben trots op jullie *powerchickies*. Vol gas voor het vak, maar ook thuis zorgen alsof je geen werk hebt. Al weet ik niet of dat laatste het beste is om na te streven;).

Lieve Lot, zelfs na een chemokuur vraag jij: hoe gaat het met je promotie? Er heeft niemand zo intens betrokken gereageerd na het zwanger worden als jij. Er is niemand die zo veerkrachtig is als jij. Ik ben zo trots op jou. En je interesse in het relatieve van dit boekje tijdens pittige tijden - ongelooflijk. Nu gewoon normaal leven oke?

Lieve koekies, lieve Beks en Daan, bij de zij-instroom wisten we natuurlijk allang dat wij de bossa's zouden worden. Huisgenoten in Deventer, Utrecht en Amsterdam. En nu alle drie bijna medisch specialist en papa's en mama. Maar veel belangrijker is hoe betrokken we altijd zijn met elkaar en elkaars families. Alle hoogte- en dieptepunten samen hebben meegemaakt. Jullie zijn voor mij heel waardevolle vrienden. En dit jaar worden we alle drie weer Antoniaan!

Lieve Kim, jij bent voor mij een voorbeeldvrouw! Zowel een topper medisch inhoudelijk, als algemeen ontwikkeld, en een top mama voor je drie boys. Het is fijn om een vrouwelijk rolmodel te hebben - die is binnen de orthopedie wat zeldzamer;). Met jou is het altijd lachen!

Luc, mijn vooropleidingmaatje, je bent voor mij het boegbeeld van het hoogst haalbare binnen het vak: extreem gedreven en ingelezen. Niet alleen in de vakliteratuur overigens. Ook de gewone boekenkast. Ik zal niet ontkennen dat jij ervoor zorgde dat ik elke keer een stap harder wilde. Superleuk dat we weer collega's worden in het UMCU!

Lieve Beltie, na onze jaren bij Bewegingswetenschappen, zijn we nu samen met Lot gewoon al 15 jaar verder (oké nog meer, maar dat wordt heel confronterend). Ik waardeer hoe attent je altijd bent, je energie om altijd iets leuks te doen, en je

ontplooiing tot echte levensgenieter inclusief danssessies in Damsco, Zwollywood en op de pompdagen. Dus Lot & Belt, op naar de volgende!

Lieve Buck, na ons fenomenale Cabaret, kan ik niet wachten om collega's te worden in het UMCU! Jij bent zo'n *powerhouse.* Ik kan veel van je leren, persoonlijk en straks ook op de OK. Maar je komt het beste tot je recht op het podium. Laten we weer eens gaan *shinen* samen!

Lieve Peter, Ellen, Tim, Jessica en Sander, al bijna 20 jaar super gastvrij, altijd ondersteunend, garant voor veel lol en liefde. Op naar nog veel meer familiemomenten!

Lieve Roland en Saskia, ik had mij geen lievere bonusouders kunnen wensen. Zo betrokken en liefdevol. En zo eigen. Ik ben heel dankbaar dat jullie mijn families compleet maken.

Lieve Bas, June, Dina, Julian, wat zijn jullie een prachtige toevoeging aan ons leven! We zijn zo dankbaar en blij dat Veda en Huub nu nichtjes, een neef en ome Basjj hebben;-)

Lieve Rosan (en Daan en Tobin), bij de NAK heb ik jou echt mogen leren kennen, ondanks dat we toen al jaren samenwoonden. Wat ben jij een top chick. En dat wij de reis van Tobin en Veda vrijwel samen mochten en mogen meemaken… daar ben ik zo dankbaar voor. Ik hoop op heel veel meer gezelligheid (en insomnia geklaag) in de toekomst.

Lieve Thijmen, Ornella en Elia, je weet pas echt dat je familie bent als je elkaar spreekt vlak na een bevalling, en de rake woorden van je stiefbroer je zo ongelooflijk helpen. Jullie zijn op afstand in Antwerpen, maar toch zo betrokken. Dank!

Lieve papa, jij bent mijn allerliefste pappie. Zo vaak heb ik jou gebeld. Altijd wat te overleggen, of discussiëren samen. Altijd breng jij rust, vertraging en reflectie. Je vindt het nooit erg als ik iets overleg, het brengt je zelden uit het lood, en je bent altijd zo'n ontzettend goed luisterend oor. Dankjewel dat je zo'n lieve papa en opa bent. Na dit boekje kan ik bij jou weer doen waar we goed in zijn: klussen!

Lieve mama, je bent mijn allerliefste mammie. Het onvoorwaardelijke; dat is zo bijzonder. Nu ik kindjes heb, realiseer ik mij meer en meer hoeveel jij voor ons hebt gedaan en om ons geeft. Mam, dankjewel, dat je altijd voor mij zorgt en dat je mij, Veda, Huub en Dennis zo veel liefhebt. Jij bent een oermoeder en geeft liefde aan Dennis en

Veda en Huubje zoals je dat aan mij en Si doet. Ik hoop dat alle volgende keren dat ik in Hotel Mama bende laptop thuisblijft. En de voorverwarmde kamers alleen maar chill-tijd betekent!

Si, zus, Bolle, schwessa, wat ben jij een TOP zus. Je bent de beste zus van de wereld. De liefste tante van de wereld. Jij brengt mij altijd weer even terug in het hier en nu. Niet haasten, niet altijd teveel willen. Jij staat garant voor lol, energie, leven, dansjes, knuffels en heel veel liefde.

Lieve Dennis, dank voor het zijn van mijn bakermat, mijn thuishaven. Ik heb heel veel aan je en waardeer je rust, creativiteit en liefde voor de kiddo's enorm. Jij hebt me altijd vrij gelaten om dit boekje te schrijven en mijn opleiding te doen, en door jou ken ik de relativiteit ervan. Ook heb je me het mooiste ooit gegeven: Veda en Huub. Ons kleine meisje nu alweer grote zus. Dit boekje heeft een wat vermoeide partner opgeleverd zo af en toe. Nu op naar nieuwe hobby's en samen avonturen maken!