BJR

SYSTEMATIC REVIEW

Patient-reported outcome measures used in patients undergoing total knee arthroplasty

A COSMIN SYSTEMATIC REVIEW

Y. Wang,
M. Yin,
S. Zhu,
X. Chen,
H. Zhou,
W. Qian

From Peking Union Medical College Hospital, Peking Union Medical College, Chinese Academy of Medical Science, Beijing, China

Patient-reported outcome measures (PROMs) are being used increasingly in total knee arthroplasty (TKA). We conducted a systematic review aimed at identifying psychometrically sound PROMs by appraising their measurement properties. Studies concerning the development and/or evaluation of the measurement properties of PROMs used in a TKA population were systematically retrieved via PubMed, Web of Science, Embase, and Scopus. Ratings for methodological quality and measurement properties were conducted according to updated COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) methodology. Of the 155 articles on 34 instruments included, nine PROMs met the minimum requirements for psychometric validation and can be recommended to use as measures of TKA outcome: Oxford Knee Score (OKS); OKS–Activity and Participation Questionnaire (OKS-APQ); 12-item short form Knee Injury and Osteoarthritis Outcome (KOOS-12); KOOS Physical function Short form (KOOS-PS); Western Ontario and McMaster Universities Arthritis Index-Total Knee Replacement function short form (WOMAC-TKR); Lower Extremity Functional Scale (LEFS); Forgotten Joint Score (FJS); Patient's Knee Implant Performance (PKIP); and University of California Los Angeles (UCLA) activity score. The pain and function subscales in WOMAC, as well as the pain, function, and quality of life subscales in KOOS, were validated psychometrically as standalone subscales instead of as whole instruments. However, none of the included PROMs have been validated for all measurement properties. Thus, further studies are still warranted to evaluate those PROMs. Use of the other 25 scales and subscales should be tempered until further studies validate their measurement properties.

Cite this article: *Bone Joint Res* 2021;10(3):203–217.

Keywords: Patient-reported outcome measure, Total knee arthroplasty, Measurement properties

Article focus

- We summarized available patient-reported outcome measures (PROMs) used in a total knee arthroplasty (TKA) population, and identified those with high quality by evaluating their measurement properties.

Key messages

- Nine instruments and six subscales in two instruments met the minimal requirements for psychometric validation, and can be recommended to use as measures for TKA outcome.
- However, none of the included PROMs have been validated for all nine measurement properties.
- Further studies are warranted to assess the measurement properties of existing instruments, especially for content validity, cross-culture validity, and structure validity.

Correspondence should be sent to Wenwei Qian; email: qianww007@163.com

doi: 10.1302/2046-3758.103.BJR-2020-0268.R1

Bone Joint Res 2021;10(3):203–217.

Strengths and limitations

- We undertook a comprehensive review under the latest COnsensus-based Standards for the selection of health Measurement Instruments (COSMIN) methodology.
- We also used meta-analysis to summarize results quantitatively.
- Our strict selection criteria meant that only a TKA population could be assessed.

Introduction

Total knee arthroplasty (TKA) has been performed for more than 40 years. It is regarded as efficacious treatment for end-stage knee arthritis, capable of improving quality of life by reducing pain and ameliorating long-term knee function.[1] Thanks to optimization of surgical methods and prosthetic designs, the ten-year survival of knee prostheses exceeds 90%.[2] However, the proportion of dissatisfied TKA recipients remains > 10%.[3,4]

TKA is an elective procedure, so patient-reported outcome measures (PROMs) are crucial to assess how well this type of intervention serves the patients' goals rather than strictly objective measures alone.[5]

PROMs have advantages over objective measurements because they: 1) largely eliminate clinicians' biases and measure health status accurately from the patients' perspective;[6] 2) enable better detection of what patients account for, and help to address possibly modifiable factors;[7] 3) aid the possibility of follow-up of patients regardless of their direct attendance; and 4) facilitate decision-making for surgical procedures.[8,9]

In clinical practice and research, it is critical (but difficult) to opt for psychometrically sound rather than frequently used PROMs for certain purposes. In this context, systematic reviews were published in 2016 and 2017 evaluating the measurement properties of PROMs in knee arthroplasty population, using COnsensus-based Standards for the selection of health Measurement Instruments (COSMIN) methodology.[10,11] Since then, many studies assessing PROMs in TKA have been conducted. Moreover, the COSMIN methodology for systematic review of PROMs has been developed further. This methodology has: 1) established specific and comprehensive guidelines for evaluating content validity;[12] 2) updated criteria for good measurement properties, risk of bias checklist,[13] approach for grading the quality of evidence, and synthesizing the overall rating; and 3) formulated recommendation standards for selection of PROMs.[14] For these reasons, we conducted a systematic review following the updated COSMIN methodology to establish a comprehensive quality assessment of PROMs for TKA.

Methods

Search strategy. We systematically searched for studies reporting the measurement properties of PROMs used in the TKA population in PubMed, Web of Science, Embase, and Scopus from commencement to 8 March 2020. To find eligible studies, we used keywords from three terms: 1) patient-reported outcome measure; 2) measurement properties (including validity, reliability, internal consistency, measurement error, responsiveness, and minimal clinically important difference (MCID)); and 3) total knee arthroplasty population. The full search strategy for PubMed is shown in detail in the Supplementary Material.

Inclusion and exclusion criteria. Articles were considered eligible if published as full texts in English and if they detailed the development or evaluated the measurement properties of PROMs used in TKA. Articles were excluded if they: 1) did not report one of the nine measurement properties or MCID; 2) did not focus on patients undergoing (or who had undergone) TKA; 3) were not published in the English language; or 4) were not full reports (e.g. only abstracts were available) because they were unlikely to contain sufficient information.

Study selection and data extraction. After removal of duplicate articles, two reviewers (SZ and XC) screened titles and abstracts independently, and identified eligible articles. Then, full manuscripts were extracted and screened for final inclusion. Discrepancies between the reviewers were resolved by discussion. The bibliographies of all selected full-text articles were screened to retrieve additional citations.

The data extraction was undertaken by two reviewers (SZ and XC) independently. First, the characteristics of included PROMs were extracted from development studies, questionnaires, and user manuals. Second, results from included studies for evaluating the methodological quality of studies and measurement properties, as well as descriptive data on feasibility and interpretability, were recorded on a standardized form.

Assessment of the measurement properties of PROMs. The measurement properties of PROMs were evaluated under the updated COSMIN methodology,[12-14] which required the following consecutive procedures: evaluation of the methodological quality and measurement properties of single studies; qualitative and quantitative summary of the results of each instrument; and grading the quality of evidence and selecting instruments.

Evaluation of the methodological quality of individual studies. The methodological quality of studies was assessed using the COSMIN risk of bias checklist, which consists of ten 'tick boxes' in accordance with development studies and nine measurement properties. Each tick box includes 3 to 35 items, which are rated as "very good", "adequate", "doubtful", or "inadequate" against standards. Two reviewers (YW and MY) completed the corresponding tick box per article independently. Besides, the overall rating of the methodological quality was based on the worst rating within each tick box.[13]

Evaluation of measurement properties for individual studies. We used the COSMIN methodology to assess content validity.[12] COSMIN set the criteria for good content validity from three aspects: 1) relevance (i.e. all items in a PROM should be relevant for the construct of interest within a specific population and context of use); 2) comprehensiveness (i.e. no key aspects of the construct should be missing); and 3) comprehensibility (i.e. the items should be understood by patients as intended). The rating for the content validity of each included PROM was conducted for a development study, content validity studies, and the instrument itself separately against the criteria. For the latter, the English version of included instruments was reviewed by English-fluent TKA-expert authors independently. Other measurement properties of the included instruments were evaluated according to the criteria described in Supplementary Table i.[14] The results of each study were rated qualitatively as "sufficient" (+), "insufficient" (–), or "indeterminate" (?). All results were evaluated by two reviewers (YW and MY) independently, and a third party (WQ) was consulted if consensus could not be reached.

Qualitative syntheses of the results of each instrument. The results for each measurement property from single studies per instrument were summarized qualitatively (i.e. + / – / ± / ?). An overall "sufficient" (+) or "insufficient" (–) rating was given if > 75% of results were concurrent. An "inconsistent" (±) rating was given if no rating exceeded 75% and no appropriate explanation for inconsistency could be given. An "indeterminate" (?) rating was given only if all single study results were indeterminate.

Quantitative syntheses of the results of each instrument. For PROMs with more than two available results on internal consistency, test–retest reliability or construct validity (i.e. Cronbach's α, intraclass correlation coefficients (ICCs), and Pearson correlation coefficients) were also pooled statistically in a meta-analysis. Weighted means and 95% confidence intervals were calculated for Pearson correlation coefficients against 36-Item Short-Form Health Survey questionnaire (SF-36; the most commonly used and validated comparator for evaluating construct validity).[15] Correlation between included scales with SF-36 pain and function subscales (convergent validity) should be higher than those with SF-36 mental and emotional subscales (divergent validity) by a minimum of 0.10. For test–retest reliability, ICCs were combined based on estimates derived from a Fisher transformation.[16] Weighted means and range of results were reported for Cronbach's α. All analyses were undertaken with use of Stata v14.0 (StataCorp, USA).

Grading the quality of evidence for each instrument. The quality of evidence was graded for each property per instrument using Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach.[14] The quality of evidence was downgraded from "high" by considering four factors: bias risk, inconsistency, imprecision, and indirectness, and eventually judged as "high", "moderate", "low", or "very low" (Supplementary Table ii).

Formulating recommendations for selection of instruments. As COSMIN suggested, each PROM was placed in a recommendation category (A to C) according to its overall ratings and quality of evidence. PROMs with sufficient content validity and at least low-quality evidence for sufficient internal consistency were placed in category A, which meant they could be recommended for use and results obtained with these PROMs could be trusted. PROMs with high-quality evidence for an insufficient measurement property were placed into category C and should not be recommended for use. PROMs were placed in category B if they were categorized neither in A nor in C, which required further research to demonstrate their measurement properties.[14]

Results

Selection and characteristics of studies. In total, 155 articles were ultimately selected from 5,145 references (Figure 1). As a result, 34 PROMs were evaluated and their characteristics are presented in Table I.[17–50] Among them, the Oxford Knee Score (OKS; 39 articles),[42] Western Ontario and McMaster Universities Arthritis Index (WOMAC; 33 articles),[48] Knee Injury and Osteoarthritis Outcome Score (KOOS; 18 articles),[25] New Knee Society Scoring System (14 articles),[32] and Forgotten Joint Score (FJS; 12 articles)[20] were the most commonly evaluated instruments, with over ten included articles.

Methodological quality and rating for measurement properties. The methodological quality and rating for content validity is displayed in Supplementary Table iii. All included studies were of "doubtful" or "inadequate" methodological quality except for Patient's Knee Implant Performance (PKIP),[45] which had an "adequate" rating for instrument development. Around half of PROMs were given an "indeterminate" rating for all three aspects of development procedures and lack of content validity studies, so content validity could be rated based only on the reviewers' rating about the instruments themselves. Supplementary Table iv shows the methodological quality and rating for the remaining measurement properties for each included study.

Overall rating and quality of evidence for included PROMs. Table II presents the qualitatively summarized ratings for the measurement properties of the included instruments. None of the PROMs reported overall ratings for all nine measurement properties, due to no rating being given for cross-cultural validity. Ratings for measurement invariance and criterion validity were reported for three and four instruments, respectively. We also analyzed quantitatively eight instruments with > 2 available results in internal consistency, reliability, or construct validity (Table III). Most results coincided except for the test–retest reliability results of Intermittent and Constant Osteoarthritis Pain (ICOAP) and Function,[23] and the sports and recreation activities subscale in the KOOS—with better rating under quantitative analyses.

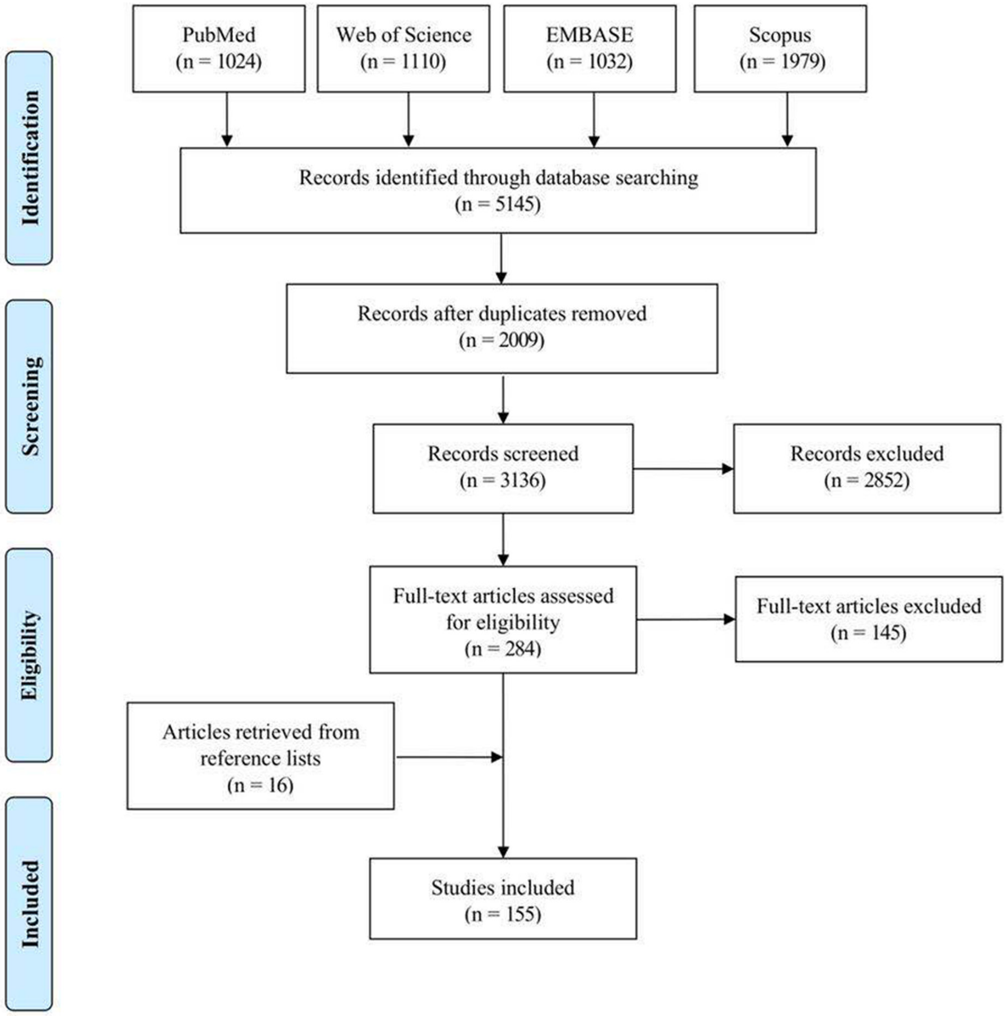

Fig. 1

Flowchart of the article selection process.

Recommendation for selection of PROMs. Eight instruments with "sufficient" ratings for the three indispensable measurement properties were placed in category A, fulfilling COSMIN standard for use as TKA outcome measures: OKS, Lower Limb Functional Scale (LEFS), KOOS Physical function Short form (KOOS-PS), FJS, WOMAC-Total Knee Replacement function short form (WOMAC-TKR), Oxford Knee Score–Activity and Participation Questionnaire (OKS-APQ), PKIP, and 12-item short form KOOS (KOOS-12).

In addition, the pain and function subscales in WOMAC, as well as the pain, function, and quality of life subscales in KOOS, were recommended for use as standalone subscales instead of whole instrument due to the "insufficient" internal consistency of the other subscales.

University of California Los Angeles (UCLA) activity score and Lower-Extremity Activity Scale (LEAS)— one-item questionnaire with sufficient content validity— also met the standard. However, we placed LEAS in category B because of "insufficient" construct validity and responsiveness.

The remaining scales were all placed in category B and required further validation studies. No included PROM was in category C.

Feasibility and interpretability. Descriptive data on feasibility and interpretability are shown in Supplementary Table v. For the recommended PROMs stated above, OKS, function subscale in WOMAC, and the function, sports and recreational activities subscale in KOOS had studies reporting > 10% items missing. Studies of ceiling effects—generally defined as > 15% of the respondents achieving the highest possible score—revealed that such effects emerged 6 to 12 months postoperatively for WOMAC, KOOS, and their adapted versions, despite the considerable variation in results between studies.

Table I. The characteristics of included patient-reported outcome measures.

Instrument	Reference to the first article	Year initially published	Construct	Recall period	Number of items	Number of domains	Response options	Scoring	Range of score	Interpretation of the score	Numbers of articles included
American Academy of Orthopaedic Surgeons (AAOS) Hip and Knee Core Scale	Johanson et al [17]	2004	Impact on quality of life of hip / knee problem	Past week	7	3	5-, 6-, 7-point ordinal response	Sum up and convert to a metric of 0 to 100	0 to 100	100 for the best possible score	3
Anterior Knee Pain Scale (AKPS)	Kujala et al [18]	1993	Anterior knee pain symptoms	Not mention	13	1	3 to 4-, 5-point ordinal response	Sum up	0 to 100	0 for the worst possible score	2
Activity Scale for Arthroplasty Patients (ASAP)	Domzalski et al[51]	2010	High level functioning and participating for arthroplasty recipients	Today	10	1	4-point Likert response	Sum up	10 to 40	10 for the worst possible score	2
Core Outcome Measures Index knee (COMI-knee)	Impellizzeri et al [19]	2016	Lack of clear description	Item dependent	6	5	5-point Likert response and NRS	Take the average of score (mean) for each domain	0 to 10	0 for the best possible score	1
Forgotten Joint Score (FJS)	Behrend et al [20]	2012	Ability to forget the artificial joint	Not mention	12	1	5-point Likert response	Sum up and convert to a metric of 0 to 100	0 to 100	0 for the worst possible score	13
High-Activity Arthroplasty Score (HAAS)	Talbot et al [21]	2010	Variation in functional ability after arthroplasty	Not mention	4	1	4-, 5-, 6-, 7-point ordinal response	Sum up	0 to 18	0 for the worst possible score	3
High-Flexion Knee Score (HFKS)	Na et al [22]	2012	Knee status in the high-function range	Not mention	9	2	5-point Likert response	Sum up	9 to 45	9 for the worst possible score	1
Intermittent and Constant Osteoarthritis Pain (ICOAP)	Hawker et al [23,24]	2007	Constant and intermittent pain experienced by hip / knee OA patients	Past week	11	2	5-point Likert response	Sum up and convert to a metric of 0 to 100 for each subscale and total score	0 to 100	0 for the best possible score	9
Knee Injury and Osteoarthritis Outcome Score (KOOS)	Roos et al [25]	1998	Symptoms and function disability in knee injury and osteoarthritis patients	Past week	42	5	5-point Likert response	Sum up and convert to a metric of 0 to 100 for each subscale	0 to 100	0 for the worst possible score	18
Knee Injury and Osteoarthritis Outcome Score Joint Arthroplasty (KOOS, JR)	Lyhman et al [26]	2016	Knee health status before and after knee arthroplasty	Past week	7	1	5-point Likert response	Sum up and convert to a metric of 0 to 100	0 to 100	0 for the worst possible score	4
Knee Injury and Osteoarthritis Outcome Physical function Short form (KOOS-PS)	Perruccio et al [27]	2008	Physical function states that represent the progression of physical disability from early to late knee OA	Past week	7	1	5-point Likert response	Sum up and convert to a metric of 0 to 100	0 to 100	0 for the best possible score	8
12-item short forms Knee Injury and Osteoarthritis Outcome (KOOS-12)	Gandek et al [28]	2018	Difficulties experienced by knee OA patients by measuring knee specific pain, function and quality of life	Past week	12	3	5-point Likert response	Sum up and convert to a metric of 0 to 100	0 to 100	0 for the worst possible score	2
Activities of Daily Living Scale of the Knee Outcome Survey (KOS-ADLS)	Irrgang et al [29]	1998	Symptoms and functional limitations experienced during activities of daily living	Past 1 to 2 days	14	2	6-point Likert response	Sum up and convert to a metric of 0% to 100%	0% to 100%	0 for the worst possible score	4
Knee Pain Questionnaire (KPQ)	Boeckstyns et al [30]	1987	Knee pain	Not mention	10	1	Dichotomous response	The sum of positive and unanswered questions	0 to 10	0 for the best possible score	2
Knee Surgery Perception Questionnaire (KSPQ)	Levinger et al [31]	2014	Discrepancy between patients' expectations and actual functional abilities preceding to knee arthroplasty surgery	Current	Part A: 20 Part B: 20	5	6-point Likert response	Part A&B: Sum up, Discrepancy score: Subtract the score of Part A from Part B	Part A&B: 0 to 120 Discrepancy score: -120 to 120	Part A&B: 0 for the worst possible score Discrepancy score: the lowest score for low expectation from surgery	1
2011 Knee Society Scoring System (KSS 2011)	Noble et al [32]	2011	Health status of the knee by evaluating pain relief, functional abilities, satisfaction, and fulfillment of expectations	Not mention	34 (7 objective, 27 subjective)	3 subjective domains	5-, 6-, 7-point ordinal response	Sum up for each subscale	Satisfaction subscale: 0 to 40 Expectation subscale: 3 to 15 Functional activities subscale: 0 to 100	The lowest score for the worst possible score for each subscale	15
Adjusted 2011 Knee Society Scoring System (KSS-A)	Dinjens et al [33]	2016	The same as KSS 2011	Not mention	25 (5 objective, 20 subjective)	5 (1 objective, 4 subjective)	3-, 5-, 6-, 7-point ordinal response and NRS	Sum up and convert to a metric of 0% to 100% for each subscale	0% to 100%	0 for the worst possible score	1
Knee Society Scoring System short form (KSS short form)	Scuderi et al [34]	2015	The same as NKSS	Not mention	10	3	3-, 5-, 6-point ordinal response and NRS	Sum up for each subscale	Lack of clear description	Lack of clear description	1

Continued

Table I. Continued

Instrument	Reference to the first article	Year initially published	Construct	Recall period	Number of items	Number of domains	Response options	Scoring	Range of score	Interpretation of the score	Numbers of articles included
Lower-Extremity Activity Scale (LEAS)	Saleh et al [35]	2005	Lower limb activity	Not mention	18	1	1 point per checked item		1 to 18	1 for the worst possible score	2
Lower Limb Functional Scale (LEFS)	Binkley et al [36]	1999	Lower limb function	Today	20	1	5-point Likert response	Sum up	0 to 80	0 for the worst possible score	3
Lequesne Algofunctional Index for the Knee (Lequesne)	Lequesne et al [37]	1987	Severity of knee OA	Not mention	11	3	2-, 3-, 5-, 7-point ordinal response	Sum up and convert to six level of severity for knee OA	0 to 24	0 for the best possible score	3
Lysholm Knee Scoring Scale (Lysholm)	Lysholm et al [38]	1982	Lack of clear description	Not mention	8	1	3-, 4-, 5-, 6-point ordinal response	Sum up	0 to 100	0 for the worst possible score	2
Modified Forgotten Joint Score (MFJS)	Robinson et al [39]	2018	Ability to forget the artificial joint	Not mention	10	1	5-point Likert response	Sum up and convert to a metric of 0% to 100%	0% to 100%	0 for the worst possible score	1
Computer-Adaptive Test for Hip and Knee OA (OA-CAT)	McDonough et al [40]	2009	Function, Pain and Disability associated with hip / knee OA	Past month	15	3	5-point ordinal response	Logit scores transformed to T-scores (Mean: 50, SD 10)			3
OsteoArthritis of Knee and Hip Quality of Life Scale (OAKHQOL)	Rat et al [41]	2006	Quality of life affected by hip / knee OA	Past 4 weeks	43	5	11-point Likert response	Sum up and convert to a metric of 0 to 100	0 to 100	0 for the worst possible score	4
Oxford Knee Score (OKS)	Dawson et al [42]	1998	Knee arthroplasty outcomes by measuring problems for knee arthroplasty recipients	Past 4 weeks	12	1	5-point Likert response	Sum up	0 to 48	0 for the worst possible score	39
Oxford Knee Score–Activity and Participation Questionnaire (OKS-APQ)	Dawson et al [43]	2014	Knee arthroplasty outcome by measuring physical activities and social participation	Past 4 weeks	8	1	5-point Likert response	Sum up and convert to a metric of 0 to 100	0 to 100	0 for the worst possible score	1
Patient Administered Questionnaires Knee (PAQ-knee)	Mancuso et al [44]	2012	Knee condition by measuring pain, function, psychological wellbeing, and satisfaction	Item dependent	29	1	4-, 5-point ordinal response and NRS	Sum up 18 of the 29 questions	0 to 100	0 for the best possible score	1
Patient's Knee Implant Performance (PKIP)	Lewis et al [45]	2014	Patients' functional performance of their arthroplasty knee.	Past week	24	4	5-, 6-, 11-point ordinal response	Sum up and convert to a metric of 0 to 100	0 to 100	0 for the worst possible score	2
Tegner Activity Scale	Tegner et al [46]	1985	Activity level	Current	10	1	1 point per checked item		0 to 10	0 for the worst possible score	3
University of California Los Angeles (UCLA) activity score	Amstutz et al [47]	1984	Activity level	Current	10	1	1 point per checked item		1 to 10	1 for the worst possible score	7
Western Ontario and McMaster Universities Arthritis Index (WOMAC)	Bellamy et al [48]	1988	Discomfort and disability in hip/knee OA patients	Past 48 hours	24	3	5-point Likert response, VAS, or NRS	Sum up for each subscale	Dependent on response opinions	0 for the best possible score	34
Western Ontario and McMaster Universities Arthritis Index-Total Knee Replacement function short form (WOMAC-TKR)	Libes et al [49]	2013	Knee specific function	Past 48 hours	7	1	VAS	Sum up and convert to a metric of 0 to 100	0 to 100	0 for the best possible score	1
Work, Osteoarthritis or joint-Replacement Questionnaire (WORQ)	Kievit et al [50]	2014	Impact on work ability of knee problems	Past week	13	1	5-point Likert response	Sum up and convert to a metric of 0 to 100	0 to 100	0 for the worst possible score	2

NRS, numerical rating scales; OA, osteoarthritis; VAS, visual analogue scale.

Table II. Overall qualitative rating and quality of evidence for measurement properties of each instrument.

Instrument	Subscales	Content validity		Structural validity		Internal consistency		Cross-cultural validity		Measurement invariance	
		Overall rating	Quality of evidence	Overall rating	Quality of evidence	Overall rating	Quality of evidence	Overall rating	Quality of evidence	Overall rating	Quality of evidence
AAOS Hip and Knee		±	Very low			?	Low				
AKPS		+	Very low			?	Low				
ASAP		±	Very low			?	Low				
COMI-knee		+	Very low								
FJS		+	Moderate	+	High	+	High				
HAAS		+	Moderate			?	Moderate				
HFKS		+	Very low								
ICOAP	Constant pain	+	Very low			?	Moderate				
	Intermittent pain	+	Very low			?	Moderate				
KOOS	Pain	+	Moderate	+	High	+	High				
	Function, daily living	+	Moderate	+	High	+	High				
	Function, sports and recreational activities	+	Moderate	+	High	+	High				
	Quality of life	+	Moderate	+	High	+	High				
	Symptoms	+	Moderate			?	Moderate				
KOOS, JR		+	Very low			?	Low				
KOOS-PS		+	Low	+	Moderate	+	High				
KOOS-12		+	Very low	+	Moderate	+	High			+	Low
KOS-ADLS		+	Very low			?	Low				
KPQ		+	Very low								
KSPQ		+	Very low								
KSS (2011)		+	Very low	?	Moderate	?	Moderate			+	Low
KSS-A		+	Very low			?	Low				
KSS short form		+	Very low	±	Moderate	?	Moderate				
LEAS		+	Very low								
LEFS		+	Very low	+	Low	+	Moderate				
Lequesne		+	Very low			?	Very low				
Lysholm		?				?	Very low				
MFJS		+	Very low			?	Low				
OA-CAT		?		+	Moderate	+	Moderate				
OAKHQOL		+	Moderate								
OKS		+	Moderate	+	High	+	High				
OKS-APQ		+	Very low	+	Low	+	Moderate			+	Very low
PAQ-knee		+	Very low			?	Low				
PKIP		+	Moderate	+	High	+	High				
Tegner		±	Very low								

Continued

Table II. Continued

Instrument	Subscales	Content validity		Structural validity		Internal consistency		Cross-cultural validity		Measurement invariance	
		Overall rating	**Quality of evidence**	**Overall rating**	**Quality of evidence**	**Overall rating**	**Quality of evidence**	**Overall rating**	**Quality of evidence**	**Overall rating**	**Quality of evidence**
UCLA		+	Very low								
WOMAC	Pain, and Function	+	Very low	+	Low	+	High				
	Stiffness	+	Very low			?	Moderate				
WOMAC-TKR		+	Very low	+	Low	+	Moderate				
WORQ		+	Moderate			?	Low				
Instrument	Subscales	Reliability	Measurement error	Criterion validity	Construct validity	Responsiveness					
		Overall rating	Quality of evidence	Overall rating	Quality of evidence	Overall rating	Quality of evidence	Overall rating	Quality of evidence	Overall rating	Quality of evidence
AAOS Hip and Knee								+	Moderate	?	Very low
AKPS								+	Moderate		
ASAP								+	Low		
COMI-knee		+	Very low	?	Very low			+	Moderate	+	Moderate
FJS		+	High	–	Low			+	High	?	Low
HAAS		+	Low	?	Low			+	Moderate		
HFKS								+	Very low	+	Very low
ICOAP	Constant pain	±	Moderate	–	Very low			+	High	+	Very low
	Intermittent pain	–	Moderate	–	Very low			+	High	+	Very low
KOOS	Pain	+	Moderate	+	Low			±	High	+	Very low
	Function, daily living	+	Moderate	+	Very low			+	High	+	Very low
	Function, sports and recreational activities	±	Moderate	–	Moderate			+	High	+	Very low
	Quality of life	+	Moderate	–	Very low			+	High	+	Very low
	Symptoms	+	Moderate	–	Low			+	High	+	Very low
KOOS, JR				+	Very low	+	Very low	+	Moderate	+	High
KOOS-PS		±	Moderate					+	High	?	Low
KOOS-12						+	Very low	+	Moderate	?	Very low
KOS-ADLS		+	Low	?	Low			+	High	+	High
KPQ		–	Very low								
KSPQ								+	Moderate		
KSS (2011)		+	High					+	Moderate	+	Low
KSS-A								+	Moderate	?	Very low
KSS short form						+	Very low	+	Moderate	?	Very low
LEAS		+	Very low					–	Moderate	–	Moderate
LEFS		+	Very low					+	Moderate	?	Very low

Continued

Table II. Continued

Instrument	Subscales	Content validity		Structural validity		Internal consistency		Cross-cultural validity		Measurement invariance	
		Overall rating	Quality of evidence	Overall rating	Quality of evidence	Overall rating	Quality of evidence	Overall rating	Quality of evidence	Overall rating	Quality of evidence
Lequesne		–	Low							?	Very low
Lysholm		+	Very low	?	Very low			+	Very low		
MFJS		+	Very low					+	Moderate		
OA-CAT										?	Very low
OAKHQOL		–	Low					–	Low	?	Low
OKS		+	High	+	Moderate			+	High	±	High
OKS-APQ		+	Very low					+	Low	?	Very low
PAQ-knee		+	Low					+	Low	+	Low
PKIP		+	Very low					+	Moderate	+	Moderate
Tegner		+	Very low	?	Very low			+	Moderate		
UCLA		+	Low	?	Very low			+	Moderate	?	Low
WOMAC	Pain, and Function	+	High	+	Low			+	Moderate	+	Low
	Stiffness	+	High	–	Low			+	Moderate	+	Low
WOMAC-TKR						+	Very low			?	Very low
WORQ		+	Moderate	+	Moderate			+	Moderate	?	Very low

+, sufficient; –, insufficient; ±, inconsistent; ?, indeterminate; AAOS Hip and Knee, AAOS Hip and Knee Core Scale; AKPS, Anterior Knee Pain Scale; ASAP, Activity Scale for Arthroplasty Patients; COMI-knee, Core Outcome Measures Index knee; FJS, Forgotten Joint Score; HAAS, High-Activity Arthroplasty Score; HFKS, High-Flexion Knee Score; ICOAP, Intermittent and Constant Osteoarthritis Pain; KOOS, Knee Injury and Osteoarthritis Outcome Score; KOOS, JR, Knee Injury and Osteoarthritis Outcome Score Joint Arthroplasty; KOOS-PS, Knee Injury and Osteoarthritis Outcome Physical function Short form; KOOS-12, 12-item short forms Knee Injury and Osteoarthritis Outcome; KOS-ADLS, Activities of Daily Living Scale of the Knee Outcome Survey; KPQ, Knee Pain Questionnaire; KSPQ, Knee Surgery Perception Questionnaire; KSS (2011), 2011 Knee Society Scoring System; KSS-A, Adjusted 2011 Knee Society Scoring System; KSS short form, Knee Society Scoring System short form; LEAS, Lower-Extremity Activity Scale; LEFS, Lower Limb Functional Scale; Lequesne, Lequesne Algofunctional Index for the Knee; Lysholm, Lysholm Knee Scoring Scale; MFJS, Modified Forgotten Joint Score; OA-CAT, Computer-Adaptive Test for Hip and Knee OA; OAKHQOL, OsteoArthritis of Knee and Hip Quality of Life Scale; OKS, Oxford Knee Score; OKS-APQ, Oxford Knee Score–Activity and Participation Questionnaire; PAQ-knee, Patient Administered Questionnaires Knee; PKIP, Patient's Knee Implant Performance; Tegner, Tegner Activity Scale; UCLA, University of California Los Angeles activity score; WOMAC, Western Ontario and McMaster Universities Arthritis Index; WOMAC-TKR, Western Ontario and McMaster Universities Arthritis Index-Total Knee Replacement function short form; WORQ, Work, Osteoarthritis or joint-Replacement Questionnaire.

Table III. Quantitative summarizing measurement properties for eligible instruments.

Instruments	Subscales	Pooled Cronbach's α Mean (Range)	Pooled ICC Mean (95% CI)	Pooled Pearson correlation coefficient Mean (95% CI)*					
				Bodily pain	Physical function	Role emotional	Mental health	PCS	MCS
FJS		0.94 (0.81, 0.97)	0.93 (0.92, 0.94)						
ICOAP	Constant pain	0.90 (0.76, 0.97)	0.76 (0.71, 0.81)						
	Intermittent pain	0.88 (0.81, 0.93)	0.73 (0.67, 0.78)						
KOOS	Pain	0.88 (0.65, 0.92)	0.88 (0.86, 0.90)	0.63 (0.63, 0.64)	0.49 (0.49, 0.49)	0.17 (0.16, 0.18)	0.27 (0.27, 0.28)		
	Function, daily living	0.94 (0.80, 0.96)	0.85 (0.82, 0.88)	0.62 (0.61, 0.62)	0.60 (0.60, 0.60)	0.29 (0.28, 0.30)	0.32 (0.32, 0.32)		
	Function, sports and recreational activities	0.88 (0.60, 0.98)	0.74 (0.70, 0.78)	0.38 (0.37, 0.38)	0.43 (0.43, 0.44)	0.12 (0.10, 0.13)	0.16 (0.16, 0.16)		
	Quality of life	0.82 (0.71, 0.92)	0.81 (0.78, 0.84)	0.53 (0.53, 0.54)	0.48 (0.47, 0.48)	0.19 (0.19, 0.20)	0.28 (0.28, 0.28)		
	Symptoms	0.75 (0.56, 0.93)	0.85 (0.83, 0.88)	0.45 (0.45, 0.45)	0.37 (0.37, 0.38)	0.17 (0.17, 0.18)	0.24 (0.24, 0.24)		
KOOS-PS		0.87 (0.79, 0.91)							
KOOS-12	Pain	0.78 (0.75, 0.84)							
	Function	0.79 (0.78, 0.82)							
	Quality of life	0.82 (0.80, 0.84)							
	Summary score	0.91 (0.90, 0.93)							
KSS (2011)	Symptoms	0.80 (0.31, 0.96)	0.89 (0.86, 0.91)	0.35 (0.32, 0.39)	0.24 (0.21, 0.26)	0.15 (0.13, 0.16)	0.17 (0.15, 0.19)	0.36 (0.34, 0.38)	0.19 (0.18, 0.20)
	Patient satisfaction	0.88 (0.79, 0.95)	0.88 (0.85, 0.90)	0.49 (0.47, 0.51)	0.32 (0.30, 0.33)	0.24 (0.22, 0.25)	0.29 (0.27, 0.30)	0.33 (0.31, 0.36)	0.30 (0.29, 0.30)
	Patient expectations	0.88 (0.73, 1.00)	0.85 (0.82, 0.87)	0.08 (0.06, 0.11)	0.13 (0.11, 0.16)	0.12 (0.12, 0.13)	0.10 (0.08, 0.11)	0.05 (0.03, 0.08)	0.10 (0.09, 0.11)
	Walking and standing	0.80 (0.68, 0.96)	0.84 (0.80, 0.87)						
	Standard activities	0.88 (0.83, 0.94)	0.86 (0.82, 0.89)						
	Advanced activities	0.84 (0.73, 0.93)	0.84 (0.80, 0.87)						
	Discretionary knee activities	0.79 (0.51, 0.94)	0.78 (0.72, 0.82)						
	Functional activities	0.90 (0.80, 0.93)	0.92 (0.91, 0.94)	0.47 (0.46, 0.49)	0.52 (0.50, 0.54)	0.28 (0.26, 0.29)	0.32 (0.30, 0.33)	0.57 (0.56, 0.59)	0.24 (0.22, 0.25)
OKS		0.92 (0.66, 0.94)	0.93 (0.92, 0.94)	0.62 (0.61, 0.63)	0.66 (0.66, 0.67)	0.32 (0.31, 0.33)	0.32 (0.31, 0.32)	0.65 (0.64, 0.66)	0.46 (0.45, 0.48)
WOMAC	Function	0.94 (0.82, 0.98)	0.91 (0.89, 0.93)	-0.56 (-0.56, –0.55)	-0.57 (-0.57, –0.57)	-0.30 (-0.31, –0.30)	-0.26 (-0.27, –0.26)		

Continued

Table III. Continued

Instruments	Subscales	Pooled Cronbach's α Mean (Range)	Pooled ICC Mean (95% CI)	Pooled Pearson correlation coefficient Mean (95% CI)*					
				Bodily pain	Physical function	Role emotional	Mental health	PCS	MCS
	Pain	0.83 (0.67, 0.91)	0.88 (0.86, 0.91)	-0.57 (-0.57, –0.56)	-0.47 (-0.48, –0.47)	-0.29 (-0.30, –0.29)	-0.28 (-0.28, –0.27)		
	Stiffness	0.84 (0.70, 0.91)	0.87 (0.84, 0.90)	-0.47 (-0.47, –0.47)	-0.36 (-0.37, –0.36)	-0.28 (-0.28, –0.27)	-0.19 (-0.19, –0.19)		

*Pearson coefficients with 36-Item Short-Form Health Survey questionnaire.

KSS (2011), 2011 Knee Society Scoring System; CI, confidence interval; FJS, Forgotten Joint Score; ICOAP, Intermittent and Constant Osteoarthritis Pain; KOOS, Knee Injury and Osteoarthritis Outcome Score; KOOS-12, 12-item short forms Knee Injury and Osteoarthritis Outcome; KOOS-PS, Knee Injury and Osteoarthritis Outcome Physical function Short form; MCS, Mental compartment summary score; OKS, Oxford Knee Score; PCS, Physical compartment summary score; WOMAC, Western Ontario and McMaster Universities Arthritis Index.

Discussion

Our review systematically summarized 155 articles evaluating the measurement properties of 34 PROMs according to the latest COSMIN methodology. Nine PROMs (OKS, LEFS, KOOS-PS, FJS, WOMAC-TKR, OKS-APQ, PKIP, KOOS-12, and UCLA) met the COSMIN standard for recommendation of use for assessing TKA outcomes. WOMAC and KOOS were recommended for use as separate subscales, rather than a total score. The other 23 instruments, including the stiffness subscale in WOMAC, and the symptoms subscale in KOOS, had the potential for use but would need further validation studies.

The strengths of our review were its size and use of the latest COSMIN methodology. We identified 34 specific PROMs in the TKA population, which exceeded the 13 scales in the review by Harris et al (five knee scores and eight lower-limb scores).[11] Thus, with new articles published in the last five years, it is no wonder that we identified more psychometrically validated PROMs than those in OKS and OKS-APQ as revealed by Harris et al.[11] With regard to WOMAC, our results concurred with the supporting measurement properties identified by Harris et al[11] for pain and function subscales, but without evidence for sufficient structure validity for the stiffness subscale. Thus, we recommended to use pain and function subscales solely.

We also identified 68 additional articles that were not included in the review by Gagnier et al.[10] Of these, 31 documented development or assessed the psychometric properties of 11 additional PROMs (Core Outcome Measures Index Knee,[19] High-Flexion Knee Score,[22] KOOS, Joint Replacement,[26,52–54] KOOS-12,[28,55] 2011 KSS,[56–64] Adjusted 2011 KSS,[33] KSS short form,[34] LEFS,[36,65,66] Modified FJS,[39] Computer-Adaptive Test for Hip and Knee OA,[40,67,68] and UCLA[40,47,66,69–72]). Eight instruments were not selected for our review: the British Orthopaedic Association Score and Original KSS were not completely patient-reported; the Physical Activity Scale for the Elderly questionnaire, Self-Efficacy for Rehabilitation outcome scale, and Short Musculoskeletal Function Assessment Questionnaire did not aim to measure TKA outcomes; Arthritis Impact Measurement Scales, Japanese Knee Osteoarthritis Measure, and the Lower Limb Activity Profile were not evaluated for their psychometric properties in a specific TKA population. We concurred with the review by Harris et al[11] as they considered Musculoskeletal Outcomes Data Evaluation and Management System and AAOS Hip and Knee Questionnaire as the same instrument, and combined their results. However, our findings are completely different from those in a systematic review by Gagnier et al[10] in that they concluded Work, Osteoarthritis or joint-Replacement Questionnaire (WORQ) to be a promising instrument. This change could be explained by the addition of recently published literature in our study, especially on those evaluating structure validity (one of the determinants for recommendation),[28,65,73–75] as well as modifications in methodological instruments. The updated COSMIN methodology established a standard for instrument selection instead of counting positive ratings for properties.

Measurement properties related to recommendation formulation. Content validity and internal consistency were related to the quality of items and internal structure of instruments. Thus, sufficient ratings of these properties constituted the basic requirement for instrument selection.[14]

Content validity. The rating for content validity of a PROM was given based on the information on development study, content validity studies, and the instrument itself.

As for PROMs development, 11/34 included development studies that provided a clear description of representative patient involvement in elicitation of items, which was in concordance with a previous study which revealed that more than one-quarter of the development procedures lacked patient involvement.[76] Approximately one-half of 11 PROMs were developed based on a TKA population, which would degrade the quality of evidence for the other half of instruments because of indirectness. PROMs were developed to reflect disease effects on patients from their perspective. Items failed to achieve this purpose if they were generated by physicians or focus groups without a targeted population. Thus, individual interviews with specific patients was the best method to enable the relevance of items.[77] In addition, no study (except the development study for PKIP) undertook and described clearly interviews with target populations about comprehensiveness and comprehensibility of PROMs.[45] This absence led to an overall "insufficient" methodological quality and "indeterminate" rating for these two aspects.

Less than half of included instruments had available content-validity studies. Of the 30 content-validity studies, five studies asked patients[78–82] and two asked experts[83,84] about relevance, comprehensiveness, or comprehensibility. The remaining studies (> 70%) were pilot studies for comprehensibility within cross-culture adaption, which was a high proportion.

Internal consistency. Cronbach's α was calculated commonly to evaluate internal consistency. However, as a prerequisite for interpreting internal consistency,[85] "sufficient" structure validity was graded only for eight instruments and six subscales in two multidimensional scales. Six studies conducted confirmatory factor analysis—regarded as the best method—to confirm the unidimensional structure of OKS, as well as the pain, function, quality of life subscales of KOOS, FJS, OKS-APQ, and PKIP.[28,43,73–75,86] Four studies undertook exploratory factor analysis (a less robust but adequate method),[32,65,87,88] but only one of them provided complete results and found the unidimensionality of LEFS.[65] We provided a "sufficient" rating for structure validity but downgraded the quality of evidence for WOMAC-TKA, KOOS-12, and KOOS-PS, as well as the pain and function subscales of WOMAC because all of their items were included in the validated subscales in KOOS.

Reliability and measurement error. Reliability and measurement error were assessed based on test–retest designs, but reliability (i.e. ICC) was more commonly calculated than measurement error. With respect to the methodological quality of test–retest designs, nearly three-quarters of studies had a "doubtful" or "inadequate" rating due to inappropriate time intervals and/or unclear test conditions. Time intervals between the two tests should be neither too short nor too long to avoid a recall bias and changes in the patient's state, respectively. A time interval of two weeks is, in general, considered appropriate, albeit this is not standardized.[89]

Hypothesis for construct validity and responsiveness. The latest COSMIN methodology deletes all standards for formulating hypotheses, and is recommended to set or adopt hypotheses by the review teams themselves.[14] This strategy led to more evidence on hypotheses for construct validity (30/34 of included instruments) and responsiveness (26/34 of included instruments). More than 80% of included studies measured responsiveness by calculating the effect size rather than a correlation coefficient. The effect size measures the magnitude of change, but gives little information about the ability of the instrument to detect changes over time (i.e. responsiveness).[89] We rated these results as "indeterminate" because we could not formulate a hypothesis without knowing the true change.

Feasibility and interpretability. In recent decades, conventional PROMs (e.g. WOMAC, KOOS, OKS) have been blamed for their postoperative ceiling effects (15% of the respondents achieving the highest possible score). We found that ceiling effects started to occur in WOMAC, KOOS, and their adapted versions six to 12 months after TKA (especially for pain subscales). Obvious ceiling effects were not revealed for OKS or other PROMs under this definition.

The ceiling effect is crucial because it prevents detection of further improvements in patients who have reached the highest score, which influences the discriminative power of the instrument. Improvements in surgical methods and changes in TKA recipients (e.g. increase in the number of younger patients with higher functional demands and expectations)[90] require expansion of the threshold for the best possible state defined 20 to 30 years previously.

As PROMs, scores for ceiling effects are thought to reflect patients' perception of their health status rather than the ability to discriminate changes.[91] The goal of TKA is to achieve freedom from pain and functional satisfaction. In addition, many patients do not experience pain postoperatively, which leads to a skewed distribution of postoperative scores.

Thus, ceiling effects have a limited impact on the use of all recommended PROMs in TKA recipients overall. However, for long-term follow-up studies (or younger patients with high activity demands), instruments with a specific construct (e.g. "forgotten artificial joints")[39] or developed for a specific population (e.g. OKS-APQ)[43] can be implemented.

Current trends. Recently, researchers have shown an increased interest in PROMs as an outcome measure for TKA. In our systematic review, though the first included PROM was introduced in 1982,[17] half of the PROMs and more than 60% of studies included were developed within the last decade. A systematic review also showed an increase in use of PROMs in TKA studies.[92]

However, the most frequently evaluated and used PROMs remain OKS, WOMAC, and KOOS, which were developed more than 20 years ago.[92] However, except for OKS, other instruments were not of sufficient quality. Also, sufficient internal consistency for subscales in KOOS and WOMAC was not demonstrated until publication of a study in 2019,[52] which highlighted the need for recommending PROMs of sufficient measurement properties to use in clinical practice and research.

In our review, no instrument was evaluated for all nine measurement properties. Also, no study of high methodological quality assessed content validity or cross-culture validity. To this extent, the recommended PROMs only met the minimal requirements for psychometric validity and still require further validation studies. Further studies are still warranted to evaluate existing PROMs, especially for a study evaluating content validity of included PROMs by interviewing patients and experts because content validity studies are the only variable determinants for this measurement property.

Additionally, we used qualitative and quantitative methods to synthesize ratings for measurement properties, and quantitative methods produced better ratings. Often, a quantitatively pooled method has a better ability to detect subtle changes than a qualitatively summarized method. Thus, we suggest using both qualitative and quantitative methods when conducting COSMIN review in the future.

Our systematic review had three main limitations. First, although we undertook exhaustive research unlikely missing any major trials, omissions might have occurred. Thus, we searched all the references of included studies manually, and included all the relevant articles. Second, we only included studies evaluating measurement properties in a TKA population. Thus, the results tested on populations combining TKA, total hip arthroplasty, or other patients were eliminated, which contributed to some PROMs and studies in previous systematic reviews not being included in our study. We believe that this strict inclusion criterion could increase the accuracy of our results in TKA patients. Finally, the recommendation formulated by our review does not necessarily mean the other 23 instruments are of "poor" quality. To some extent, they require further robust studies to evaluate their measurement properties.

In conclusion, nine PROMs and six subscales in two PROMs met the minimum requirements for psychometric validation and can be recommended for use as measures of TKA outcome. These are OKS, LEFS, KOOS-PS, FJS, WOMAC-TKR, OKS-APQ, PKIP, KOOS-12, UCLA, the pain and function subscales in WOMAC, and the pain, function, and quality-of-life subscales in KOOS. However, none of the included PROMs have been validated for all measurement properties. Thus, further studies are still warranted to evaluate those PROMs. Use of the other 25 scales and subscales should be tempered until further studies validate their measurement properties.

Supplementary material

Includes PubMed search strategy, tables showing COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) criteria and approach, and tables documenting methodological quality, qualitative rating, and descriptive data for each instrument per article. References of all included studies are also provided.

References

1. **Price AJ, Alvand A, Troelsen A, et al.** Knee replacement. *The Lancet.* 2018;392(10158):1672–1682.
2. **O'Connor MI.** Implant survival, knee function, and pain relief after TKA: are there differences between men and women? *Clin Orthop Relat Res.* 2011;469(7):1846–1851.
3. **Halawi MJ, Jongbloed W, Baron S, Savoy L, Williams VJ, Cote MP.** Patient Dissatisfaction After Primary Total Joint Arthroplasty: The Patient Perspective. *J Arthroplasty.* 2019;34(6):1093–1096.
4. **Nam D, Nunley RM, Barrack RL.** Patient dissatisfaction following total knee replacement: a growing concern? *Bone Joint J.* 2014;96-B(11 Supple A):96–100.
5. **Franklin PD, Li W, Ayers DC.** The Chitranjan Ranawat Award: functional outcome after total knee replacement varies with patient attributes. *Clin Orthop Relat Res.* 2008;466(11):2597–2604.
6. **Hamilton DF, Giesinger JM, Giesinger K.** It is merely subjective opinion that patient-reported outcome measures are not objective tools. *Bone Joint Res.* 2017;6(12):665–666.
7. **Gunaratne R, Pratt DN, Banda J, Fick DP, Khan RJK, Robertson BW.** Patient Dissatisfaction Following Total Knee Arthroplasty: A Systematic Review of the Literature. *J Arthroplasty.* 2017;32(12):3854–3860.
8. **Dakin H, Eibich P, Beard D, Gray A, Price A.** The use of patient-reported outcome measures to guide referral for hip and knee arthroplasty. *Bone Joint J.* 2020;102-B(7):950–958.
9. **Price AJ, Kang S, Cook JA, et al.** The use of patient-reported outcome measures to guide referral for hip and knee arthroplasty. *Bone Joint J.* 2020;102-B(7):941–949.
10. **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;32(5):1688–1697.
11. **Harris K, Dawson J, Gibbons E, et al.** Systematic review of measurement properties of patient-reported outcome measures used in patients undergoing hip and knee arthroplasty. *Patient Relat Outcome Meas.* 2016;7:101–108.
12. **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.
13. **Mokkink LB, de Vet HCW, Prinsen CAC, et al.** COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. *Qual Life Res.* 2018;27(5):1171–1179.
14. **Prinsen CAC, Mokkink LB, Bouter LM, et al.** COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual Life Res.* 2018;27(5):1147–1157.
15. **Ware JE Jr, Sherbourne CD.** The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care.* 1992;30(6):473–483.
16. **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. *Osteoarthritis and Cartilage.* 2016;24(8):1317–1329.

17. **Johanson NA**, **Liang MH**, **Daltroy L**, **Rudicel S**, **Richmond J**. American Academy of orthopaedic surgeons lower limb outcomes assessment instruments. reliability, validity, and sensitivity to change. *J Bone Joint Surg Am*. 2004;86-A(5):902–909.
18. **Kujala UM**, **Jaakkola LH**, **Koskinen SK**, **Taimela S**, **Hurme M**, **Nelimarkka O**. Scoring of patellofemoral disorders. *Arthroscopy*. 1993;9(2):159–163.
19. **Impellizzeri FM**, **Leunig M**, **Preiss S**, **Guggi T**, **Mannion AF**. The use of the core outcome measures index (COMI) in patients undergoing total knee replacement. *Knee*. 2017;24(2):372–379.
20. **Behrend H**, **Giesinger K**, **Giesinger JM**, **Kuster MS**. The "forgotten joint" as the ultimate goal in joint arthroplasty: validation of a new patient-reported outcome measure. *J Arthroplasty*. 2012;27(3):430–436.
21. **Talbot S**, **Hooper G**, **Stokes A**, **Zordan R**. Use of a new high-activity arthroplasty score to assess function of young patients with total hip or knee arthroplasty. *J Arthroplasty*. 2010;25(2):268–273.
22. **Na S-E**, **Ha C-W**, **Lee C-H**, **SE N**, . A new high-flexion knee scoring system to eliminate the ceiling effect. *Clin Orthop Relat Res*. 2012;470(2):584–593.
23. **Hawker GA**, **Davis AM**, **French MR**, **et al**. Development and preliminary psychometric testing of a new OA pain measure--an OARSI/OMERACT initiative. *Osteoarthritis Cartilage*. 2008;16(4):409–414.
24. **Hawker GA**, **Stewart L**, **French MR**, **et al**. Understanding the pain experience in hip and knee osteoarthritis – an OARSI/OMERACT initiative. *Osteoarthritis Cartilage*. 2008;16(4):415–422.
25. **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.
26. **Lyman S**, **Lee Y-Y**, **Franklin PD**, **Li W**, **Cross MB**, **Padgett DE**. Validation of the KOOS, jr: a short-form knee arthroplasty outcomes survey. *Clin Orthop Relat Res*. 2016;474(6):1461–1471.
27. **Perruccio AV**, **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. *Osteoarthritis Cartilage*. 2008;16(5):542–550.
28. **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). *Osteoarthritis Cartilage*. 2019;27(5):746–753.
29. **Irrgang JJ**, **Snyder-Mackler L**, **Wainner RS**, **Fu FH**, **Harner CD**. Development of a patient-reported measure of function of the knee. *J Bone Joint Surg Am*. 1998;80-A(8):1132–1145.
30. **Boeckstyns MEH**. Development and construct validity of a knee pain questionnaire. *Pain*. 1987;31(1):47–52.
31. **Levinger P**, **Diamond NT**, **Menz HB**, **et al**. Development and validation of a questionnaire assessing discrepancy between patients' pre-surgery expectations and abilities and post-surgical outcomes following knee replacement surgery. *Knee Surg Sports Traumatol Arthrosc*. 2016;24(10):3359–3368.
32. **Noble PC**, **Scuderi GR**, **Brekke AC**, **et al**. Development of a new Knee Society scoring system. *Clin Orthop Relat Res*. 2012;470(1):20–32.
33. **Dinjens RN**, **Grimm B**, **Heyligers IC**, **Senden R**. Adjustments in 2011 KSS increase the clinical suitability. *Acta Orthop Belg*. 2016;82(1):43–51.
34. **Scuderi GR**, **Sikorskii A**, **Bourne RB**, **Lonner JH**, **Benjamin JB**, **Noble PC**. The Knee Society Short Form Reduces Respondent Burden in the Assessment of Patient-reported Outcomes. *Clin Orthop Relat Res*. 2016;474(1):134–142.
35. **Saleh KJ**, **Mulhall KJ**, **Bershadsky B**, **et al**. Development and validation of a lower-extremity activity scale. use for patients treated with revision total knee arthroplasty. *J Bone Joint Surg Am*. 2005;87-A(9):1985–1994.
36. **Binkley JM**, **Stratford PW**, **Lott SA**, **Riddle DL**. The lower extremity functional scale (LEFS): scale development, measurement properties, and clinical application. North American orthopaedic rehabilitation research network. *Phys Ther*. 1999;79-4:371–383.
37. **Lequesne MG**, **Mery C**, **Samson M**, **Gerard P**. Indexes of severity for osteoarthritis of the hip and knee. Validation-value in comparison with other assessment tests. *Scand J Rheumatol Suppl*. 1987;65:85–89.
38. **Lysholm J**, **Gillquist J**. Evaluation of knee ligament surgery results with special emphasis on use of a scoring scale. *Am J Sports Med*. 1982;10(3):150–154.
39. **Robinson PG**, **Rankin CS**, **Lavery J**, **Anthony I**, **Blyth M**, **Jones B**. The validity and reliability of the modified forgotten joint score. *J Orthop*. 2018;15(2):480–485.
40. **McDonough CM**, **Stoiber E**, **Tomek IM**, **et al**. Sensitivity to change of a computer adaptive testing instrument for outcome measurement after hip and knee arthroplasty and periacetabular osteotomy. *J Orthop Sports Phys Ther*. 2016;46(9):756–767.
41. **Rat A-C**, **Pouchot J**, **Coste J**, **et al**. Development and testing of a specific quality-of-life questionnaire for knee and hip osteoarthritis: OAKHQOL (OsteoArthritis of Knee Hip Quality Of Life). *Joint Bone Spine*. 2006;73(6):697–704.
42. **Dawson J**, **Fitzpatrick R**, **Murray D**, **Carr A**. Questionnaire on the perceptions of patients about total knee replacement. *J Bone Joint Surg Br*. 1998;80-B(1):63–69.
43. **Dawson J**, **Beard DJ**, **McKibbin H**, **Harris K**, **Jenkinson C**, **Price AJ**. Development of a patient-reported outcome measure of activity and participation (the OKS-APQ) to supplement the Oxford knee score. *Bone Joint J*. 2014;96-B(3):332–338.
44. **Mancuso CA**, **Ranawat AS**, **Meftah M**, **Koob TW**, **Ranawat CS**. Properties of the patient administered questionnaires: new scales measuring physical and psychological symptoms of hip and knee disorders. *J Arthroplasty*. 2012;27(4):575–582.
45. **Lewis S**, **Price M**, **Dwyer KA**, **et al**. Development of a scale to assess performance following primary total knee arthroplasty. *Value in Health*. 2014;17(4):350–359.
46. **Tegner Y**, **Lysholm J**. Rating systems in the evaluation of knee ligament injuries. *Clin Orthop Relat Res*. 1985;198:43–49.
47. **Amstutz HC**, **Thomas BJ**, **Jinnah R**, **Kim W**, **Grogan T**, **Yale C**. Treatment of primary osteoarthritis of the hip. A comparison of total joint and surface replacement arthroplasty. *J Bone Joint Surg Am*. 1984;66-A(2):228–241.
48. **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;5(2):231–241.
49. **Liebs TR**, **Herzberg W**, **Gluth J**, **et al**. Using the patient's perspective to develop function short forms specific to total hip and knee replacement based on WOMAC function items. *Bone Joint J*. 2013;95-B(2):239–243.
50. **Kievit AJ**, **Kuijer PPFM**, **Kievit RA**, **Sierevelt IN**, **Blankevoort L**, **Frings-Dresen MHW**. A reliable, valid and responsive questionnaire to score the impact of knee complaints on work following total knee arthroplasty: the WORQ. *J Arthroplasty*. 2014;29(6):1169–1175.
51. **Domzalski T**, **Cook C**, **Attarian DE**, **Kelley SS**, **Bolognesi MP**, **Vail TP**. Activity scale for arthroplasty patients after total hip arthroplasty. *J Arthroplasty*. 2010;25(1):152–157.
52. **Lyman S**, **Lee Y-Y**, **McLawhorn AS**, **Islam W**, **MacLean CH**. What are the minimal and substantial improvements in the HOOS and KOOS and jr versions after total joint replacement? *Clin Orthop Relat Res*. 2018;476(12):2432–2441.
53. **Austin DC**, **Torchia MT**, **Werth PM**, **Lucas AP**, **Moschetti WE**, **Jevsevar DS**. A One-Question Patient-Reported Outcome Measure Is Comparable to Multiple-Question Measures in Total Knee Arthroplasty Patients. *J Arthroplasty*. 2019;34(12):2937–2943.
54. **Khalil LS**, **Darrith B**, **Franovic S**, **Davis JJ**, **Weir RM**, **Banka TR**. Patient-Reported Outcomes Measurement Information System (PROMIS) Global Health Short Forms Demonstrate Responsiveness in Patients Undergoing Knee Arthroplasty. *J Arthroplasty*. 2020;36(6):1540–1544.
55. **Gandek B**, **Roos EM**, **Franklin PD**, **Ware JE**. A 12-Item short form of the knee injury and osteoarthritis outcome score (KOOS-12): tests of reliability, validity and responsiveness. *Osteoarthritis Cartilage*. 2019;27(5):762–770.
56. **Liu D**, **He X**, **Zheng W**, **et al**. Translation and validation of the simplified Chinese new Knee Society Scoring System. *BMC Musculoskelet Disord*. 2015;16(1):391.
57. **Hamamoto Y**, **Ito H**, **Furu M**, **et al**. Cross-cultural adaptation and validation of the Japanese version of the new Knee Society Scoring System for osteoarthritic knee with total knee arthroplasty. *J Orthop Sci*. 2015;20(5):849–853.
58. **Maniar RN**, **Maniar PR**, **Chanda D**, **Gajbhare D**, **Chouhan T**. What is the Responsiveness and Respondent Burden of the New Knee Society Score? *Clin Orthop Relat Res*. 2017;475(9):2218–2227.
59. **Kim SJ**, **Basur MS**, **Park CK**, **et al**. Crosscultural Adaptation and Validation of the Korean Version of the New Knee Society Knee Scoring System. *Clin Orthop Relat Res*. 2017;475(6):1629–1639.
60. **Silva ALPE**, **Croci AT**, **Gobbi RG**, **Hinckel BB**, **Pecora JR**, **Demange MK**. Translation and validation of the new version of the Knee Society Score - The 2011 KS Score - into Brazilian Portuguese. *Rev Bras Ortop*. 2017;52(4):506–510.
61. **Culliton SE**, **Bryant DM**, **MacDonald SJ**, **Hibbert KM**, **Chesworth BM**. Validity and Internal Consistency of the New Knee Society Knee Scoring System. *Clin Orthop Relat Res*. 2018;476(1):77–84.
62. **Kayaalp ME**, **Keller T**, **Fitz W**, **Scuderi GR**, **Becker R**. Translation and Validation of the German New Knee Society Scoring System. *Clin Orthop Relat Res*. 2019;477(2):383–393.
63. **Özden F**, **Tuğay N**, **Umut TB**, **Yalın KC**. Psychometrical properties of the Turkish translation of the new knee Society scoring system. *Acta Orthop Traumatol Turc*. 2019;53(3):184–188.

64. **Nishitani K**, **Yamamoto Y**, **Furu M**, **et al**. The minimum clinically important difference for the Japanese version of the new Knee Society Score (2011KSS) after total knee arthroplasty. *J Orthop Sci.* 2019;24(6):1053–1057.

65. **Naal FD**, **Impellizzeri FM**, **Torka S**, **Wellauer V**, **Leunig M**, **von Eisenhart-Rothe R**. The German lower extremity functional scale (LEFS) is reliable, valid and responsive in patients undergoing hip or knee replacement. *Qual Life Res.* 2015;24(2):405–410.

66. **Naal FD**, **Impellizzeri FM**, **Lenze U**, **Wellauer V**, **von Eisenhart-Rothe R**, **Leunig M**. Clinical improvement and satisfaction after total joint replacement: a prospective 12-month evaluation on the patients' perspective. *Qual Life Res.* 2015;24(12):2917–2925.

67. **Jette AM**, **McDonough CM**, **Haley SM**, **et al**. A computer-adaptive disability instrument for lower extremity osteoarthritis research demonstrated promising breadth, precision, and reliability. *J Clin Epidemiol.* 2009;62(8):807–815.

68. **Jette AM**, **McDonough CM**, **Ni P**, **et al**. A functional difficulty and functional pain instrument for hip and knee osteoarthritis. *Arthritis Res Ther.* 2009;11(4):R107.

69. **Naal FD**, **Impellizzeri FM**, **Leunig M**. Which is the best activity rating scale for patients undergoing total joint arthroplasty? *Clin Orthop Relat Res.* 2009;467(4):958–965.

70. **SooHoo NF**, **Li Z**, **Chenok KE**, **Bozic KJ**. Responsiveness of patient reported outcome measures in total joint arthroplasty patients. *J Arthroplasty.* 2015;30(2):176–191.

71. **Ghomrawi HM**, **Lee YY**, **Herrero C**, **et al**. A Crosswalk Between UCLA and Lower Extremity Activity Scales. *Clin Orthop Relat Res.* 2017;475(2):542–548.

72. **Cao S**, **Liu N**, **Li L**, **Lv H**, **Chen Y**, **Qian Q**. Simplified Chinese version of University of California at Los Angeles activity score for arthroplasty and arthroscopy: cross-cultural adaptation and validation. *J Arthroplasty.* 2017;32(9):2706–2711.

73. **Coles T**, **Williams V**, **Dwyer K**, **Mordin M**. Psychometric Evaluation of the Patient's Knee Implant Performance Questionnaire. *Value Health.* 2018;21(11):1305–1312.

74. **Martín-Fernández J**, **García-Maroto R**, **Sánchez-Jiménez FJ**, **et al**. Validation of the Spanish version of the Oxford knee score and assessment of its utility to characterize quality of life of patients suffering from knee osteoarthritis: a multicentric study. *Health Qual Life Outcomes.* 2017;15(1):15–1.

75. **Hamilton DF**, **Loth FL**, **Giesinger JM**, **et al**. Validation of the English language Forgotten Joint Score-12 as an outcome measure for total hip and knee arthroplasty in a British population. *Bone Joint J.* 2017;99-B(2):218–224.

76. **Wiering B**, **de Boer D**, **Delnoij D**. Patient involvement in the development of patient-reported outcome measures: a scoping review. *Health Expect.* 2017;20(1):11–23.

77. **Rat A-C**, **Pouchot J**, **Guillemin F**, **et al**. Content of quality-of-life instruments is affected by item-generation methods. *Int J Qual Health Care.* 2007;19(6):390–398.

78. **Heijbel S**, **Naili JE**, **Hedin A**, **W-Dahl A**, **Nilsson KG**, **Hedström M**. The Forgotten Joint Score-12 in Swedish patients undergoing knee arthroplasty: a validation study with the Knee Injury and Osteoarthritis Outcome Score (KOOS) as comparator. *Acta Orthop.* 2020;91(1):88–93.

79. **Paradowski PT**, **Kęska R**, **Witoński D**. Validation of the Polish version of the Knee injury and Osteoarthritis Outcome Score (KOOS) in patients with osteoarthritis undergoing total knee replacement. *BMJ Open.* 2015;5(7):e6947.

80. **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;20(6):1330–1341.

81. **Monticone M**, **Capone A**, **Frigau L**, **et al**. Development of the Italian version of the high-activity arthroplasty score (HAAS-I) following hip and knee total arthroplasty: cross-cultural adaptation, reliability, validity and sensitivity to change. *J Orthop Surg Res.* 2018;13(1):81.

82. **Xie F**, **Li S-C**, **Lo NN**, **et al**. Cross-cultural adaptation and validation of Singapore English and Chinese Versions of the Oxford Knee Score (OKS) in knee osteoarthritis patients undergoing total knee replacement. *Osteoarthritis Cartilage.* 2007;15(9):1019–1024.

83. **Cao S**, **Liu N**, **Han W**, **et al**. Simplified Chinese version of the forgotten joint score (FJS) for patients who underwent joint arthroplasty: cross-cultural adaptation and validation. *J Orthop Surg Res.* 2017;12(1):12–1.

84. **Gonzalez Sáenz de Tejada M**, **Escobar A**, **Herdman M**, **Herrera C**, **García L**, **Sarasqueta C**. Adaptation and validation of the osteoarthritis knee and hip quality of life (OAKHQOL) questionnaire for use in patients with osteoarthritis in Spain. *Clin Rheumatol.* 2011;30(12):1563–1575.

85. **Tavakol M**, **Dennick R**. Making sense of Cronbach's alpha. *Int J Med Educ.* 2011;2:53–55.

86. **Harris K**, **Dawson J**, **Doll H**, **et al**. Can pain and function be distinguished in the Oxford Knee Score in a meaningful way? An exploratory and confirmatory factor analysis. *Qual Life Res.* 2013;22(9):2561–2568.

87. **Thomsen MG**, **Latifi R**, **Kallemose T**, **Barfod KW**, **Husted H**, **Troelsen A**. Good validity and reliability of the forgotten joint score in evaluating the outcome of total knee arthroplasty. *Acta Orthop.* 2016;87(3):280–285.

88. **Swanenburg J**, **Koch PP**, **Meier N**, **Wirth B**. Function and activity in patients with knee arthroplasty: validity and reliability of a German version of the Lysholm score and the Tegner activity scale. *Swiss Med Wkly.* 2014;144:w13976.

89. **De Vet HCW**, **Terwee CB**, **Mokkink LB**, **Knol DL**. *Measurement in medicine: a practical guide.* 338. Cambridge: Cambridge University Press, 2011. https://www.cambridge.org/core/terms

90. **Pabinger C**, **Lothaller H**, **Geissler A**. Utilization rates of knee-arthroplasty in OECD countries. *Osteoarthritis Cartilage.* 2015;23(10):1664–1673.

91. **Marx RG**, **Jones EC**, **Atwan NC**, **Closkey RF**, **Salvati EA**, **Sculco TP**. Measuring improvement following total hip and knee arthroplasty using patient-based measures of outcome. *J Bone Joint Surg Am.* 2005;87-A(9):1999–2005.

92. **Siljander MP**, **McQuivey KS**, **Fahs AM**, **Galasso LA**, **Serdahely KJ**, **Karadsheh MS**. Current trends in patient-reported outcome measures in total joint arthroplasty: a study of 4 major orthopaedic journals. *J Arthroplasty.* 2018;33(11):3416–3421.

Author information:

- Y. Wang, MD, Surgeon
- S. Zhu, MD, Surgeon
- X. Chen, MD, Surgeon
- W. Qian, MD, Chief surgeon
  Department of Orthopedic Surgery, Peking Union Medical College Hospital, Peking Union Medical College, Chinese Academy of Medical Science, Beijing, China.
- M. Yin, BSc, Researcher, Institute of Basic Medical Sciences Chinese Academy of Medical Sciences, School of Basic Medicine Peking Union Medical College, Beijing, China.
- H. Zhou, MD, Surgeon, Plastic Surgery Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China.

Author contributions:

- Y. Wang: Designed the study, Performed the systematic evaluation, Drafted the article.
- M. Yin: Performed the systematic evaluation and statistical analysis.
- S. Zhu: Acquired and extracted the data.
- X. Chen: Acquired and extracted the data.
- H. Zhou: Drafted the manuscript.
- W. Qian: Advised and revised the study design, Edited the manuscript.

Funding statement:

- This study was supported by grants from the Peking Union Medical College Hospital Foundation (No. ZC201904183), Non-profit Central Research Institute Fund of Chinese Academy of Medical Science (No. 2019XK320010), and the National Natural Science Foundation of China (No. 81972046).No benefits in any form have been received or will be received from a commercial party related directly or indirectly to the subject of this article.

ICMJE COI statement:

- Each author certifies that he or she has no commercial associations (e.g. consultancies, stock ownership, equity interest, patent/licensing arrangements, etc.) that might pose a conflict of interest in connection with the submitted article.

Acknowledgements:

- We thank Professor Wang Li from the Department of Epidemiology and Biostatistics, Institution of Basic Medical Sciences, Peking Union Medical College, for her advice on the article design and statistical analysis.