BMJ Open

Minimal clinically important change and difference for knee osteoarthritis outcome measurement tools after non-surgical interventions: a systematic review

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2022-063026
Article Type:	Original research
Date Submitted by the Author:	30-Mar-2022
Complete List of Authors:	Silva, M Denika; University of Canberra Perriman, Diana; College of Medicine and Health Sciences, Australian National University; University of Canberra Fearon, A; University of Canberra Couldrick, Jacqui; University of Canberra Scarvell, Jennie; University of Canberra; Trauma and Orthopaedic Research Unit, Canberra Hospital
Keywords:	STATISTICS & RESEARCH METHODS, Knee < ORTHOPAEDIC & TRAUMA SURGERY, Musculoskeletal disorders < ORTHOPAEDIC & TRAUMA SURGERY

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Minimal clinically important change and difference for knee osteoarthritis outcome measurement tools after non-surgical interventions: a systematic review

M Denika C Silva[1,2,5], Diana M Perriman[1,2,3], Angela M Fearon[1,2,4], Jacqui M Couldrick[1,2], Jennie M Scarvell[1,2]

1. Faculty of Health, University of Canberra, Australia

2. Trauma and Orthopaedic Research Unit, Canberra Hospital, Canberra, Australia

3. College of Medicine and Health Sciences, Australian National University, Canberra, Australia

4. Research Institute for Sport and Exercise, University of Canberra, Australia

5. Department of Physiotherapy, General Sir John Kotelawala Defence University, Werahera, Sri Lanka

Corresponding author: M Denika C Silva (ORCID ID:0000-00018648-471X)

Address: Faculty of Health, University of Canberra, Bruce ACT 2617 Australia

Tel.: +61 (0) 433857153 **Email:** denika.silva@canberra.edu.au

Minimal clinically important change and difference for knee osteoarthritis outcome measurement tools after non-surgical interventions: a systematic review

ABSTRACT

Objectives: To systematically review and provide estimates of the minimal clinically important change (MCIC) and difference (MCID) for outcome tools in people with knee osteoarthritis (OA) after non-surgical interventions.

Design: A systematic review.

Data sources: MEDLINE, CINAHL, Web of Science, Scopus, and Cochrane databases were searched up to September 21, 2021.

Study selection and data synthesis: We included studies that calculated MCIC and MCID using any calculation method including anchor, consensus and distribution methods, for any knee OA outcome tool after non-surgical interventions. We used quality assessment tools appropriate to the studies' methods to screen out low-quality studies. Values were pooled and expressed as median and range, for each method.

Results: Forty-eight studies were eligible (anchor-k=12, consensus-k=1 and distribution-k=35). MCIC values for 13 outcome tools including Knee injury and Osteoarthritis Outcome Score (KOOS)-pain, activities of daily living (ADL), quality of life (QOL) and Western Ontario and McMaster Universities Arthritis Index (WOMAC)-function were estimated using five high-quality anchor studies. MCID values for 23 tools including KOOS-pain, ADL, QOL and WOMAC-function, stiffness and total were estimated using six high-quality anchor studies. One moderate quality consensus study reported MCIC for pain, function and global assessment. Minimum detectable

change values (MDC, from distribution method estimates) for 126 tools including KOOS-QOL and WOMAC-total were estimated using 38 good-to-fair-quality studies.

Conclusion: Median MCIC, MCID and MDC estimates were reported for outcome tools in people with knee OA after non-surgical interventions. Clinicians and researchers may consider these MCIC and MCID values when designing or interpreting clinical trials.

PROSPERO registration number: CRD42020153962

Keywords: minimal clinically important change; minimal clinically important difference; minimum detectable change; knee osteoarthritis

Strengths and limitations of this study

- We estimated MCIC (within-group), MCID (between-groups), and minimum detectable change values using anchor, consensus, or distribution methods papers respectively.
- This systematic review included a defined population of people with knee OA, after non-surgical interventions.
- High-quality anchor studies were used to contribute to MCIC and MCID estimates were assessed using a credibility tool specially designed to evaluate anchor method papers.
- Consensus and distribution methods papers were evaluated using quality assessment tools suited to each method.
- Median estimates were used to reflect the pooled data due to data skewness.

INTRODUCTION

The efficacy of therapeutic interventions is commonly evaluated using statistical significance regardless of clinical importance[1]. To understand whether differences in outcome measures after treatment are clinically important, it is necessary to know what constitutes a minimum important change or difference for the individual or cohort. These changes and differences are called the minimal clinically important change (MCIC) and difference (MCID). There are numerous outcome measures for knee osteoarthritis (OA) and many estimates of MCIC and MCID. However, these estimates can arise from different methodologies leading to variability, confusion and misinterpretation[2-5]. Achieving clarity in this space is crucial as these values are used in regulatory and clinical decision making[6,7]. This systematic review aimed to provide estimates of MCIC and MCID for knee OA outcome measurement tools in people with knee OA after non-surgical interventions.

MCIC and MCID are defined as the minimum value of an outcome measure that the patient, clinician or relevant others perceive as an important change or difference[4,8,9]. The MCIC is defined as the change in a clinical outcome measure within a single group or an individual over time. In contrast, MCID describes the difference between independent groups (for example, control versus treatment groups) or between individuals[4,10-12]. However, the terminology of MCIC and MCID is used inconsistently[13]. The concept was first described by Jaeschke, who studied patients' perceptions of pre- and post-intervention beneficial change[8]. This concept later included both improvement[14] and worsening[15]. Other commonly used terms are "minimal important change" for MCIC and "minimal important difference" for MCID.

Three methods are used to estimate MCIC and MCID: anchor, consensus and distribution[6,16]. For the anchor method, MCIC or MCID values are estimated using patients' responses where values of the outcome measure are related to an externally validated scale or 'anchor'[17]. The "global rating of change" is commonly used as the anchor in this method. The receiver operating characteristic (ROC) method for deriving an estimate from anchor questions has been suggested to be more precise for clinical settings than the mean change method[9,15]. In the consensus method, values are directly estimated by a group of experienced clinicians or patients until a consensus is achieved[6]. In the distribution method, values are estimated statistically, based on the variance of the outcome data using half the standard deviation (SD)[18], one standard error of measurement (SEM)[19] or minimum detectable change (MDC) which is based on SEM[6,20]. The anchor method is widely considered to be the most valid because it is associated with a definition of minimal importance and is based on patient perception[16,21].

Knee OA is a common cause of pain and disability[22]. Outcome measurement tools that include the domains of pain, physical function, patient global assessment and imaging are recommended to determine the efficacy of therapeutic interventions in knee OA studies[23]. MCIC and MCID values have been estimated for knee OA outcome measures in these domains using anchor, consensus, and distribution methods with variable results. The variability of the methods used makes the selection of an appropriate estimate confusing for clinicians, researchers and regulatory bodies[2-4].

The primary objective of this systematic review was to estimate MCIC and MCID for knee OA outcome measurement tools based on pooled estimates from high-quality anchor studies only. Secondary objectives were to determine MCIC and MCID estimates based on consensus and distribution methods including MDC outcomes.

METHODS

This systematic review was designed and reported according to the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) statement[24]. The protocol was registered on PROSPERO (registration number: CRD42020153962).

Literature search

Five databases (MEDLINE, CINAHL, Web of Science, Scopus and Cochrane) were searched from each database's respective inception up to September 21, 2021. A comprehensive search strategy was developed to capture all relevant articles, and database-specific MESH terms were used. The search strategy was as follows. (*knee OR genu OR tibiofemoral OR patellofemoral) AND (osteoarthr* OR degenerat*) AND (("MCIC" OR "MCID" OR "MCII" OR "MIC" OR "MII" OR "MPCC" OR "MPCD" OR "MPCI" OR "MDC" OR "SDC" OR "SDD" OR "CIC" OR "CID") OR ("minim* clinical* important change*" OR "minim* clinical* important difference*" OR "minim* clinical* important improvement*" OR "minim* important change*" OR "minim* important difference*" OR "minim* important improvement*" OR "minim* perceptible clinical* change*" OR "minim* perceptible clinical* difference*" OR "minim* perceptible clinical* improvement*" OR "minim* detectable change*" OR "small* detectable change* " OR

"small* detectable difference* " OR "clinical* important change*" OR "clinical* important difference*")). The records were exported to EndNote version X9.2 for reference management.

Study screening and selection criteria

Covidence software (Covidence systematic review software, Veritas Health Innovation, Melbourne, Australia (www.covidence.org) was used to manage the selection process. Records identified in the search were uploaded and duplicates were removed. Screening of titles and abstracts, then full texts, were performed independently by two reviewers (DS and JC) and conflicts were resolved by a third reviewer (JS). Included studies incorporated any design that calculated MCIC and MCID for any knee OA outcome measurement tool considering improvement after non-surgical intervention for adults with knee OA, and using any calculation method: anchor, consensus or distribution methods. We included studies that reported MDC because MDC is considered as an estimate from the distribution method[3]. We considered studies with MCIC or MCID values for improvement only and excluded values for deterioration because improvement values are used to evaluate the efficacy of treatment. Studies were excluded if the data from participants with knee OA could not be separated from other conditions, e.g. hip OA or other knee pathologies. Studies of MCIC, MCID and MDC were included even if they used a different terminology e.g. minimal important change for MCIC, minimal important difference for MCID and smallest detectable change or difference or minimal detectable difference for MDC.

For consistency, we defined the MCIC as the pre-post change of one group i.e. threshold for those who responded that they had minimally improved. The MCID was

defined as the difference (pre-post-change) between two groups i.e. “minimally improved” and “stayed the same” groups using the anchor response as defined in the previous studies[10,12,25]. The MDC is the minimum change above the measurement error based upon a given level of confidence[6,20]. MDC values for a 90 or 95 confidence interval are labelled as MDC90 or MDC95.

Quality assessment

The quality of the included studies was assessed according to their methodology. The quality of the anchor studies was assessed using the credibility instrument developed by Devji (2020)[26] which was specially designed to assess the anchor studies. This instrument includes five core criteria (the anchor is rated by the patient, the anchor is interpretable and relevant to the patient, the MCIC or MCID estimate is precise, the correlation between the anchor and the outcome measure reported by the patient is satisfactory, and the authors select a threshold on the anchor that reflects a small but important difference). We considered the paper to be “high” quality if at least three of the five criteria were “yes”, “definitely yes” or “to a great extent”; and of “low” quality if not[27]. Consensus studies were assessed using the Critical Appraisal Screening Program (CASP)-qualitative tool[28] which is designed to assess qualitative studies and is well-suited to consensus studies. The quality was rated as “high”, “moderate” or “low” based on reliability and credibility[29]. Distribution studies were evaluated using the National Heart, Lung and Blood Institute, National Institute of Health (NHLBI, NIH) quality assessment tool for before-after (pre-post) studies with no control group[30] and ratings included “good”, “fair” or “poor” based on reliability and credibility[30]. The quality assessment of included studies was performed

by one reviewer (DS) and a random sample of 20% had an independent second review (AF or JC) to improve the accuracy[31].

Data extraction and analysis

We extracted study characteristics including sample size, participant demographics, details of the intervention, follow-up time, outcome measurement tools, calculation method and actual estimate reported based on the method (MCIC or MCID or MDC). Additionally, we extracted the details of the anchor used in each study.

We extracted reported MCIC, MCID and MDC values from each study. We normalised the values to a 0 to 100 scale. If a study reported MDC as a percentage of the grand mean (MDC divided by grand mean percentage)[32], we converted the data into MDC90 or 95 for the pooled synthesis.

The median and range (minimum and maximum) of MCIC, MCID and MDC were calculated using multiple estimates from the included studies arising from different non-surgical interventions, calculation methods, time points, and anchors. Mean values were not calculated due to skewness of distributions[33]. We excluded low-quality anchor studies from the median MCIC and MCID synthesis. Furthermore, we conducted a sub-analysis to determine median MCID based on the ROC method where available because the ROC estimates are considered to be more precise than mean-change estimates and recommended at both individual and group level analyses, and in clinical settings[9,15]. Though we planned to conduct a sub-analysis to determine the effect of follow-up time on MCIC and MCID, we were unable to do reliable rate estimates because of a limited number of studies. Therefore, we plotted

the values against time including only studies where the outcome measures were assessed at three-time points or more.

Patient and public involvement

This is a systematic review. Patients or the public were not involved in this study.

RESULTS

Study selection

The search yielded 2376 studies and after duplicates were removed, 1059 records were screened. Two hundred and seventeen studies were screened in full-text review resulting in 48 eligible studies (k=48) **(Figure 1)**. No further studies were identified after checking the reference lists of included studies.

Included studies calculated MCIC and MCID by anchor method (k=12), by consensus method (k=1) and MDC by distribution method (k=35).

The methodological quality of included studies

Most anchor studies (k=10) were of high quality and two were of low quality[34,35] **(Supplement 1)**. The quality of the consensus study (k=1) was moderate **(Supplement 2)**. The quality of distribution studies ranged from good (k=11) to fair (k=24) **(Supplement 3)**.

Study characteristics of included studies

All the anchor studies were observational prospective cohort studies[14,15,34,36-42] and two of them were nested within randomised controlled trials[35,43] **(Table 1)**. The number of participants in each study ranged from 41 to 1606. The mean age and body mass index ranged from 57.1 to 67.9 years and from 28.1 to 33 kg/m^2, respectively. The interventions used in these studies were rehabilitation, exercise, physiotherapy and non-steroidal anti-inflammatory drugs. The follow-up time ranged from seven days to one year. Eleven studies used global rating of change as the external anchor while one study[43] used multiple anchors. Most anchor studies (k=6)[35-37,38,41,43] reported MCID values only, four studies[14,39,40,42] reported MCIC values only and two studies[15,34,] reported both MCIC and MCID. Four of these studies[34,36,40,41] also reported MDC values. Moreover, studies used different anchor questions and group classifications when calculating MCIC and MCID. For example, one MCIC study[34] considered the minimal improved response group as “my knee has got better” and another study[39] considered the response group as both “good” and “excellent” improvement groups **(Supplement 4)**.

Table 1: Characteristics of included studies

Study	Number of participants; Age, years, Mean (SD)	Intervention (Follow-up)	Outcome tool	Calculation method (Measure extracted)
Angst, 2018[36]	190; 66.1 (10.2)	Rehabilitation (3 months)	WOMAC-pain, function, functional standing/walking and stiffness; SF-36-bodily pain, physical functioning, role-physical, vitality, social functioning, mental and general	Anchor (MCID); Distribution (MDC95)
Harris, 2013[34]	134; 59 (11)	Non-operative (3 months)	OKS- pain, function, summary; ICOAP; KOOS-Pain	Anchor (MCIC, MCID); Distribution (MDC90)
Hmamouchi, 2012[37]	173; 57.1 (10.1)	NSAID (6 weeks)	WOMAC-total	Anchor (MCID)
Klokker, 2016[35]	41; NR	exercises (12 weeks)	DAP	Anchor (MCID)
Lee, 2017[43]	165; 61	Thai Chi or Physical therapy (12 weeks)	PROMIS-Short Forms- physical function, pain interference, depression, anxiety	Anchor (MCID)
Mills, 2016[15]	272; NR	Non-surgical (26,52 weeks)	KOOS-pain, activities of daily living, quality of life	Anchor (MCIC and MCID)
Mostafaee, 2021[38]	142; 58.3 (7.6)	Physiotherapy (4 weeks)	KOOS-pain, symptoms, activities of daily living, sports and recreation, and quality of life	Anchor (MCID)
Ornetti, 2011[39]	881; 67.1 (10.9)	NSAID (4 weeks)	WOMAC-function; Patient-reported functional disability on an NRS; Physician reported functional disability on an NRS	Anchor (MCIC)
Perrot, 2013[42]	1606; 66.9 (9.0)	Usual care (7 days)	Pain on a 0-10 NRS at rest and movement	Anchor (MCIC)
Singh, 2014[40]	137; NR	Conservative (2 weeks)	ICOAP-pain, constant pain, intermittent pain; KOOS-Pain; KOOS-quality of life	Anchor (MCIC); Distribution (MDC90)
Tubach, 2005[14]	603; 67.9 (10.2)	NSAID (4 weeks)	Pain on VAS on movement; Patient's global assessment of disease activity on a VAS; WOMAC-function subscale	Anchor (MCIC)
Williams, 2012[41]	159; NR	Exercise (2, 6, 12 months)	KOS-activities of daily living subscale; LEFS; WOMAC-total	Anchor (MCID); Distribution (MDC90, MDC95)
Salottolo, 2018[44]	27 (clinicians); NR	NR	Pain, Function and Global assessment	Consensus (MCIC)
Alghadir, 2015[45]	65; 54.9 (9.9)	NR	Timed Up and Go test	Distribution (MDC95)
Alghadir, 2016 a[46]	121; 52.9 (9.3)	NR (48 hours)	WOMAC- Arabic version-pain, function, total	Distribution (MDC95)
Alghadir, 2016 b[47]	121; 54.0 (9.3)	NR (48 hours)	NRS for pain- Arabic version	Distribution (MDC95)
Alghadir, 2017[48]	97; 58 (11.5)	NR (1 week)	OKS- Arabic version	Distribution (MDC95)
Alghadir, 2018[49]	121; 52.9 (12.5)	NR (24 hours)	VAS for pain; NRS for pain; Verbal Rating Scale for pain	Distribution (MDC95)
Baert, 2018[51]	8 (12 knees); 68 (11)	NR (3 seconds)	Knee joint position sense test using analogue inclinometer; Knee force sense test using a handheld dynamometer	Distribution (MDC95)
Brisson, 2018[52]	46; 61.0 (6.6)	NR (6 months, 24 months)	Peak KAM, KAM impulse and Peak KFM using 3D motion analysis; Quadriceps strength and power using dynamometer; Load frequency using a triaxial accelerometer; BMI	Distribution (MDC95)
Callghan, 2009[53]	55; 64 (14)	NR (24- 72 hours)	Quadriceps fatigue using surface electromyography	Distribution (MDC95)
Hoglund, 2019[55]	20; 58.3 (8.05)	NR (2-7 days)	30-second fast-paced walk test	Distribution (MDC95)
Hunter, 2006[77]	72; 58.9 (8.6)	Novel ther (6 months)	Cartilage volume using X-ray, MRI (Femur, Patella, Tibia)	Distribution (MDC95)
Ijima, 2019[56]	59; 59.1 (6.1)	NR (1 month)	Stopwatch-based 11- Step stair climb test	Distribution (MDC90, MDC95)

Jansen, 2021[57]	103; NR	NR (1 month)	Knee image: bone density, joint space width, osteophytes, eminence height, and joint angle	Distribution (MDC95)
Kanko, 2019[75]	74; 57.7 (8.8)	Exercises (1 week)	Star excursion balance test	Distribution (MDC95)
Kean, 2010[58]	20; 53.6 (9.1)	NR (same day)	Quadriceps isokinetic strength, isometric strength and percent of voluntary activation using a dynamometer	Distribution (MDC90)
Klokker, 2015[59]	20; 64 (6.6)	NR (1 week)	DAP	Distribution (MDC95)
McCarthy, 2004[32]	15; 65.1 (11.3)	Exercises (1 week)	Aggregated locomotor function score; Eight-meter walk time; Stair ascent and descent time; Transferring time	Distribution (MDC95)
McCarthy, 2008[60]	55; 64.9 (9.7)	NR (1 week)	Maximum isometric strength; Vastus medialis oblique, Vastus lateralis, Rectus femoris-initial median frequency; Vastus medialis oblique, Vastus lateralis, Rectus femoris fatigue slope	Distribution (MDC95)
Monticone, 2021[61]	102; 69.1 (9.0)	NR (10 days)	Fremantle Knee Awareness Questionnaire, Italian version	Distribution (MDC95)
Motyl, 2013[78]	15; 61.0 (7.8)	Steroid injection (20 days)	20-meter walk test	Distribution (MDC95)
Mutlu, 2015[76]	73 (141 knees); 56.2 (6.7)	Physiotherapy (4 weeks)	Pressure pain threshold using a handheld pressure algometer	Distribution (MDC90)
Nalbant, 2021[62]	25; 62.3 (9.8)	NR (same day)	L-test	Distribution (MDC95)
Naylor, 2014[63]	75; 67.6 (9.4)	NR (10 days)	VAS for pain; Six Minute Walk Test; Timed Up and Go test; KOOS-pain, symptoms, activities of daily living, quality of life	Distribution (MDC95)
Parveen, 2017[64]	25; 50.5 (6.3)	NR (7 days)	Tinetti Performance-Oriented Mobility Assessment scale-total, balance subscale, gait subscale	Distribution (MDC95)
Peter, 2018[65]	135; NR	NR (1 week)	Animated activity questionnaire	Distribution (MDC95)
Piva, 2004[66]	25; 57 (8)	NR (same day)	Get up and Go test	Distribution (MDC95)
Pratheep, 2018[67]	20; 70.3 (5.8)	NR (2 weeks)	Pressure pain threshold using algometer	Distribution (MDC95)
Ravaud, 1999[74]	30; 65.4 (7.7)	NR (2 weeks)	Joint space width	Distribution (MDC95)
Suhail and Chaudha, 2021[546]	82; 60.4 (6.7)	NR (48 hours)	2-minute walk test	Distribution (MDC95)
Suwit, 2020[502]	55; 69.0 (11)	NR (1 week)	30-seconds chair stand; 40-meter fast-paced test; 9-step climb	Distribution (MDC90)
Takacs, 2014[68]	20; 64.1 (7.9)	NR (14 days)	Single-leg standing balance test	Distribution (MDC95)
Tevald, 2016[69]	25; 61 (9)	NR (7 days)	Hip abductor strength using a hand-held dynamometer	Distribution (MDC95)
Tse, 2021[70]	38; 65.1 (7.0)	NR (same day)	Frontal plane tibial alignment using inclinometer	Distribution (MDC95)
Turcot, 2008[71]	25; 63.9 (7.6)	NR (8 days)	3D linear accelerations of the tibia and femur during walking	Distribution (MDC95)
Van der Straaten, 2020[72]	19; 65.1 (5.2)	NR (20 days)	Range of motion (trunk abduction-adduction, pelvis abduction-adduction, hip flexion-extension) and Center of mass displacement	Distribution (MDC95)
Yuruk, 2014[73]	40; NR	NR (One day)	De Motion Mobility Index (Turkish version)	Distribution (MDC90, MDC95)

MCIC: Minimal Clinical Important Change, MCID: Minimal Clinically Important Difference, MDC90: Minimum Detectable Change based on 90% confidence interval, MDC95: Minimum Detectable Change based on 95% confidence interval, NR: Not Reported, WOMAC: Western Ontario and McMaster Universities Arthritis Index, SF-36: 36-item Short Form health survey, OKS: Oxford Knee Score, ICOAP: Intermittent and Constant Osteoarthritis Pain, KOOS-PS: Knee injury and Osteoarthritis Outcome Score Physical Function Short form, KOOS: Knee injury and Osteoarthritis Outcome Score, DAP: Dynamic weight-bearing Assessment of Pain, PROMIS: Patient-Reported Outcome Measurement Information System, NRS: numeric rating scale, VAS: Visual Analog Scale, KOS: Knee Outcome Survey, LEFS: Lower Extremity Functional Scale, KAM: knee adduction moment, 3-dimensional: 3D, KFM: knee flexion moment, MRI: Magnetic Resonance Imaging

The consensus study (k=1)[44] **(Table 1)** used a questionnaire to survey 27 clinicians from a range of specialities (orthopaedic (38%), rheumatology (33%), internal medicine (19%), and other (9%)). The clinicians were asked about MCIC values for pain, function and global assessment for severe knee OA. However, participants were not asked to consider time duration nor the interventions. MCIC was termed "minimal clinically important improvement" in this study.

Most distribution studies (k=30) were test-retest observational studies[45-74] assessing the reliability of the outcome tool, and five used datasets from interventional cohort studies[75,76] and randomised controlled trials[32,77,78] **(Table 1)**. In the distribution studies (k=35), the number of participants in each study ranged from 8 to 135. The mean age and body mass index ranged from 50.5 to 70.3 years and from 22.7 to 35 kg/m^2, respectively. Studies estimated MDC90 and MDC95. The follow-up time ranged from the same day to one year.

The MCIC estimates derived using the anchor method

The median MCIC for 13 tools (with subscales) were calculated based on five high-quality anchor studies[14,15,39,40,42] using 23 estimates (**Table 2**). These estimates were based on different underlying calculations, follow-up time and anchor questions. Methods for calculating MCIC included: mean change (pre and post mean change of the minimally improved group)[8,15,40], 75th centile value of the mean change of the group[39,42] and 75th centile value adjusted with the baseline score, age, and disease duration[14]. Most studies included one follow-up time (range: 7 days to 4 weeks), but one study[15] reported MCIC at two time-points (26 and 52 weeks). One anchor question was used in most studies, however, two studies[15,39] reported different MCIC values

based on two different anchor questions (general health status and functional state). All data were pooled for the median estimates regardless of differences in calculation methods, follow-up time, and anchor questions.

Table 2: Minimal Clinical Important Change (MCIC) values of knee osteoarthritis outcome tools derived using the anchor method

Outcome tools	Score	Median	Minimum	Maximum	Number of estimates	Study
ICOAP-pain	100=worst	**18.5**			1	Singh, 2014[40]
ICOAP-constant pain	100=worst	**18.7**			1	Singh, 2014[40]
ICOAP-intermittent pain	100=worst	**18.4**			1	Singh, 2014[40]
KOOS-pain	100=best	**12.4**	4.3	20.1	4	Mills, 2016[15]
KOOS-activities of daily living	100=best	**8.4**	8.2	8.7	2	Mills, 2016[15]
KOOS-quality of life	100=best	**9.8**	8.0	11.6	2	Mills, 2016[15]; Singh, 2014[40]
KOOS-PS	100=worst	**2.2**			1	Singh, 2014[40]
Pain on NRS at rest	100=worst	**10.0**	10	10	2	Perrot, 2013[42]
Pain on VAS on movement	100=worst	**19.9 mm**			1	Tubach, 2005[14]
Patient-reported functional disability on an NRS	100=worst	**27.6**	27.2	27.9	2	Ornetti, 2011[39]
Patient's global assessment of disease activity on VAS	100=worst	**18.3 mm**			1	Tubach, 2005[14]
Physician-reported functional disability on an NRS	100=worst	**25.3**	25	25.5	2	Ornetti, 2011[39]
WOMAC-function	100=worst	**17.0**	9.1	17.1	3	Tubach, 2005[14]; Ornetti, 2011[39]
Estimates based on low-quality studies						
ICOAP-pain	100=worst	**13.4**			1	Harris, 2013[34]
KOOS-PS	100=worst	**12.0**			1	Harris, 2013[34]
OKS-pain	100=best	**17.3**			1	Harris, 2013[34]
OKS-function	100=best	**10.6**			1	Harris, 2013[34]
OKS-summary	100=best	**7.1**			1	Harris, 2013[34]

MCIC: Minimal Clinical Important Change, ICOAP: Intermittent and constant osteoarthritis pain, KOOS: Knee injury and Osteoarthritis Outcome Score, KOOS-PS: Knee injury and Osteoarthritis Outcome Score Physical Function Short form, NRS: numeric rating scale, VAS: Visual Analog Scale, WOMAC: Western Ontario and McMaster Universities Arthritis Index, OKS: Oxford Knee Score

All the scores are from 0 to 100

The median estimates are shaded in blue
Estimates based on low-quality studies are shaded in grey

The MCID estimates derived using the anchor method

The median MCID for 23 tools were calculated based on six high-quality anchor studies[15,36-38,41,43,45] using 83 estimates (**Table 3**). These estimates were based on different underlying calculations, follow-up time, and anchor questions. Methods for calculating MCID included: ROC method[79] only (k=3)[37,38,41], mean change (Redilmier and Lorig) method only (pre-post mean change difference between two groups)[80] (k=1)[36], both ROC and mean change methods (k=1)[15] , and mean change in T-scores with multiple anchors (k=1)[43]. Most studies (k=4) included one follow-up (range: 4 to 12 weeks), but two studies included MCID at multiple time-points e.g. 2, 6 and 12 months[15,41]. One anchor question was used in most studies, however, one study used multiple anchors[43]. All data were pooled for the median estimates regardless of differences in calculation methods, follow-up time, and anchor questions.

Table 3: Minimal Clinical Important Difference (MCID) values of knee osteoarthritis outcome tools derived using the anchor method

Outcome tools	Score	Median	Minimum	Maximum	Number of estimates	Study
KOOS-pain	100=best	**11.8**	4	17	13	Mills, 2016[15]; Mostafaee, 2021[38]
KOOS-activities of daily living	100=best	**2.5**	-1.5	15.5	7	Mills, 2016[15]; Mostafaee, 2021[38]
KOOS-quality of life	100=best	**6.5**	3	12.5	7	Mills, 2016[15]; Mostafaee, 2021[38]
KOOS-sports/recreation	100=best	**17.5**			1	Mostafaee, 2021[38]
KOOS-symptoms	100=best	**12.5**			1	Mostafaee, 2021[38]
KOS-activities of daily living	100=best	**6.4**	2.2	10.6	6	Williams, 2012[41]
LEFS	100=best	**6.9**	0.6	15.6	6	Williams, 2012[41]
PROMIS Short Forms-physical function*	100=worst		4.5	5.3	NR	Lee, 2017[43]
PROMIS Short Forms-pain interference*	100=worst		4.2	4.2	NR	Lee, 2017[43]
PROMIS Short Forms-depression*	100=worst		6	6.2	NR	Lee, 2017[43]
PROMIS Short Forms-anxiety*	100=worst		4.5	6.6	NR	Lee, 2017[43]
WOMAC-pain	100=best	**8.7**	7.1	21	3	Angst, 2018[36]
WOMAC-function	100=best	**14.5**	11.3	14.9	3	Angst, 2018[36]
WOMAC-stiffness	100=best	**20.2**	16.2	23.8	3	Angst, 2018[36]
WOMAC-standing/walking	100=best	**8.1**	5.9	10.2	2	Angst, 2018[36]
WOMAC-total	100=best	**6.8**	1.6	16.8	8	Williams, 2012[41]; Hmamouchi, 2012[37]
SF-36-bodily pain	100=best	**9.3**	8.2	10.4	2	Angst, 2018[36]
SF-36-physical function	100=best	**4.0**	3.8	4.2	2	Angst, 2018[36]
SF-36-role-physical	100=best	**8.5**	7.5	9.5	2	Angst, 2018[36]
SF-36-vitality	100=best	**4.5**	4.16	4.9	2	Angst, 2018[36]
SF-36-social function	100=best	**4.8**	2.6	7.0	2	Angst, 2018[36]
SF-36-mental health	100=best	**4.1**	2.9	5.2	2	Angst, 2018[36]
SF-36-general health	100=best	**6.6**	6	7.2	2	Angst, 2018[36]
Estimates based on low-quality studies						
DAP	100=worst	**24**			1	Klokker, 2016[35]
ICOAP-pain	100=worst	**7.8**			1	Harris, 2013[34]
KOOS-PS	100=worst	**7.8**			1	Harris, 2013[34]
OKS-pain	100=best	**14.3**			1	Harris, 2013[34]
OKS-function	100=best	**9.5**			1	Harris, 2013[34]
OKS-summary	100=best	**6.4**			1	Harris, 2013[34]

MCID: Minimal Clinically Important Difference, OA: osteoarthritis, KOOS: Knee injury and Osteoarthritis Outcome Score, KOS: Knee Outcome Survey, LEFS: Lower Extremity Functional Scale, PROMIS: Patient-Reported Outcome Measurement Information SystemSF36: 36-item Short Form health survey, WOMAC: Western Ontario and McMaster Universities Arthritis Index, DAP: Dynamic weight-bearing Assessment of Pain, ICOAP: Intermittent and constant osteoarthritis pain,KOOS-PS: Knee injury and Osteoarthritis Outcome Score Physical Function Short form, OKS: Oxford Knee Score

All the scores are from 0 to 100
*Estimates were reported as a range
The median estimates are shaded in blue
Estimates based on low-quality studies are shaded in grey

MCID estimates based on the ROC method were reported and compared with pooled MCID estimates for all methods. Four[15,37,38,41] of six studies (67%) used the ROC method. The ROC estimates were the same as the overall pooled estimates in most cases **(Table 4).**

Table 4: Comparison of receiver operating curve (ROC) method based Minimal Clinical Important Difference (MCID) estimates with overall pooled estimates

Outcome tools	ROC method-based estimates, Median (range)	ROC method-based, Number of estimates (study)	Pooled estimates regardless of calculation method, Median (range)	Number of estimates- Pooled regardless of calculation method, (study)
KOOS-pain	**11.8 (4.0 to 18.0)**	8 (Mills, 2016[15]; Mostafaee, 2021[38])	11.8 (4.0 to 17.0)	13 (Mills, 2016[15]; Mostafaee, 2021[38])
KOOS-activities of daily living	**12 (-1.5 to 155)**	5 (Mills, 2016[15]; Mostafaee, 2021[38])	12.5 (-1.5 to 15.5)	7 (Mills, 2016[15]; Mostafaee, 2021[38])
KOOS-quality of life	**6.5 (3.0 to 12.5)**	6 (Mills, 2016[15]; Mostafaee, 2021)	6.5 (3.0 to 12.5)	7 (Mills, 2016[15]; Mostafaee, 2021[38])
KOOS-sports/recreation	**17.5**	1 (Mostafaee, 2021[38])	17.5	1Mostafaee, 2021[38])
KOOS-symptoms	**12.5**	1 (Mostafaee, 2021[38])	12.5	1 (Mostafaee, 2021[38])
KOS-ADL	**6.4 (2.2 to 10.6)**	6 (Williams, 2012[41])	6.4 (2.2 to 10.6)	6 (Williams, 2012[41])
LEFS	**6.9 (0.6 to 15.6)**	6 (Williams, 2012[41])	6.9 (0.6 to 15.6)	6 (Williams, 2012[41])
WOMAC-total	**7.8 (1.6 to 16.8)**	8 (Williams, 2012[43]; Hmamouchi, 2012[37])	7.8 (1.6 to 16.8)	8 (Williams, 2012[41]; Hmamouchi, 2012[37];)

ROC: Receiver Operating Curve, MCID: Minimal Clinically Important Difference, KOOS: Knee injury and Osteoarthritis Outcome Score, KOS: Knee Outcome Survey, LEFS: Lower Extremity Functional Scale, WOMAC: Western Ontario and McMaster Universities Arthritis Index

All the estimates are out of 100.

The median estimates are shaded in blue
Shaded in green: pooled estimates of these MCID values were based on the ROC method only

The effect of follow-up time on MCIC and MCID

There were insufficient data to establish reliable rate estimates for the effect of time. The MCIC of Knee injury and Osteoarthritis Outcome Score (KOOS)-pain and KOOS-quality of life (QOL) and, the MCID of KOOS-pain, KOOS-QOL, Knee Outcome

Score (KOS)-activity of daily living (ADL), Lower Extremity Functional Scale (LEFS) and Western Ontario and McMaster Universities Arthritis Index (WOMAC)-total were assessed at more than three different time points. The MCIC of KOOS-QOL and, MCID of KOOS-QOL, LEFS and KOS-ADL appeared to increase with increasing follow-up time. However, MCIC of KOOS-pain and MCID of WOMAC-total appeared to reduce with follow-up time and KOOS-pain remained constant (**Supplement 5).**

MCIC values derived using the consensus method

One consensus study[44] reported that MCIC for pain, function and global assessment were 20% of the maximum score.

The MDC estimates derived using the distribution method

The median MDC was calculated for 126 tools based on 38 studies (35 good-to-fair distribution and three high-quality anchor studies) using 308 estimates (**Supplement 6)**. These estimates were based on different calculation methods and follow-up times. Four included studies reported MDC90 values only[40,50,58,76] and 29 studies MDC95 values only. Five studies[36,41,56,63,73] reported both the MDC90 and MDC95 values. Most studies (k=37) reported unadjusted MDC, while one study reported both the adjusted and unadjusted estimates[36]. Six studies separately reported inter/intra-rater MDC values[45,51,55,59,66,72]. Furthermore, three studies reported distinct values for two patient groups in each study, for example, the placebo group and the treatment group[77], the most painful and the least painful groups[69], and the groups that reported moderate improvement ("great deal better") and MCID improvement ("somewhat better")[40]. Most studies assessed the index (worst) knee, but one study based the estimate on all diseased knees (12 knees in 8 patients)[51].

Regarding the time point of MDC estimation, most studies (k=39) reported MDC estimates at one time-point only, but one study reported MDC estimates at three-time points (2, 6 and 12 months)[41]. All data were pooled for the median estimates regardless of differences in calculation methods and follow-up time.

DISCUSSION

This systematic review provided estimates for MCIC and MCID of knee OA outcome tools after non-surgical interventions derived using anchor, consensus and distribution methods respectively. This review is unique in that it provides pooled estimates for MCIC, MCID (based on high-quality studies) and MDC (from good to fair-quality studies) of knee OA outcome tools after non-surgical interventions. MDC was reported for a greater number of outcome measures (126) than for MCIC (13) or MCID (23). MCID estimates based on the ROC method were similar to the overall pooled median estimates, however, the majority of MCID studies used the ROC method. Although we found that some MCIC and MCID appear to increase with follow-up time, this was not consistent.

The estimates for MCIC and MCID reported in this review are lower than those reported previously[81-83]. Previous reviews which included knee replacement interventions[81-83] produced higher estimates suggesting that knee replacement cohorts need more improvement to be satisfied. The MCID values for WOMAC-pain and function in this review ranged from 7.1 to 21 and 11.3 to 14.9 (out of 100) respectively; compared to reviews of total knee replacements which reported values ranging from 4.0 to 47.9 and 1.8 to 33.0 (out of 100)[81-83]. This disparity may be due to

differences in disease severity which has been previously reported based on baseline pain score[6,15,84]. Therefore, our data is more applicable to patients and cohorts receiving non-surgical interventions. Furthermore, previous knee OA intervention studies have used MCID estimates from studies with combined hip and knee OA[85-86]. Given that MCID is sensitive to disease type[6], our median estimates are likely to be more applicable to the knee OA population.

MDC was reported for more outcome measures than MCIC or MCID. MDC is derived from data distribution only, unlike MCIC and MCID which are related to patients' perception[17,21]. Researchers may use MDC estimates as an option if MCIC or MCID are not reported. Yet, according to the results of this study and others, MDC can be larger or smaller than MCID[4,10]. Hence, researchers using MDC estimates from single studies to establish a sample size may over-or under-estimate the number of participants required for a given power.

The ROC estimates were similar to the pooled MCID estimates. Our pooled median estimates are based on a combination of the mean change[80] and ROC methods[79]. There is strong support for the exclusive use of ROC because it can be applied to both individuals and groups and is recommended in clinical settings[9,15]. Moreover, the area under the curve of the ROC has the advantage of being able to interpret the level of confidence for the MCID estimate from acceptable to outstanding discrimination between responders and non-responders[17]. Therefore, we recommend using our median ROC based MCID estimates where possible.

Although we found that some MCIC and MCID appear to increase with follow-up time, this was not consistent. Two previous studies[15,41] suggested that there may be an effect of time, due to changing of perceptions over time (response-shift), especially in patients with chronic conditions[15,87]. Additionally, recall bias is affected by increased follow-up time and may also affect estimates[6,26,88]. Therefore, although the consistency of follow-up time must be considered, more data is required to determine the effect of follow up time on MCIC and MCID.

One of the studies included in this review used the consensus method. They reported that MCIC was 20% of the maximum score for pain, function and global assessment[44], but, our anchor studies data suggest that MCIC is highly variable (2.2 to 27.6 out of 100) depending on the outcome measurement. Therefore, the blanket application of 20% may not be suggested regardless of the tool used.

In this review, we considered only MCIC, MCID and MDC but there are other measures of clinical improvement. While the MCIC and MCID are used to assess meaningful clinical effects, recent reports have questioned the applicability of these concepts as they do not consider the costs, risks, benefits and inconvenience of the treatment. The smallest worthwhile effect (SWE) was developed using the benefit-harm trade-off method, described by Barrett in 2005[89]. The SWE is defined as the smallest amount of improvement which is identified by the patient as worthwhile when considering the improvement outweighing risks and inconvenience[90] and the estimates are always compared to natural recovery[91]. However, only one study has reported SWE for people undergoing total knee replacement[92]. Other studies have evaluated “patients acceptable symptom state” (PASS) which is the symptom state

that patients consider acceptable or when they feel "well" after treatment[83,93,94]. PASS estimates for WOMAC function are reported to be between 31 and 34.4[95]. These values are much higher than our MCIC median estimate of 17 (9.1 to 17.1). Although MCIC and MCID are still commonly used, the development of this field of research will enable value judgements as well as clinical judgements to be considered in the interpretation of clinical trials of interventions.

This systematic review should be considered in light of its limitations. Median estimates were used to reflect the pooled data due to data skewness, but some of the ranges were wide, challenging the certainty of some of these estimates. Previous evidence suggests that data from follow-up times of less than one month are more reliable[26,96,97]. However, we included all estimates regardless of follow-up time. Statistical analysis was not conducted to determine the effect of follow-up time due to limited available data, but this is an interesting area for further study. The certainty of evidence using Grading of Recommendations Assessment, Development and Evaluation (GRADE) was not attempted due to the heterogeneity of the data. The grey literature was not searched for this review.

This review presents median estimates for MCIC, MCID and MDC for people with knee OA following non-surgical interventions. A subset of MCID estimates is reported based on the ROC method. The pooled ROC data are recommended for both individual and group analyses and clinical settings. MDC estimates are available for more outcome measures but are purely statistical and arguably less applicable. This review clarifies the current understanding of MCIC, MCID and MDC in the knee OA population.

Author contributions

Conception and design: M Denika C Silva, Diana M Perriman, Angela M Fearon, Jennie M Scarvell

Data acquisition: M Denika C Silva, Jacqui M Couldrick

Data-analysis and Interpretation of the data: M Denika C Silva, Diana M Perriman, Angela M Fearon, Jennie M Scarvell

Drafting of the article revising it critically for important intellectual content: M Denika C Silva, Diana M Perriman, Angela M Fearon, Jacqui M Couldrick Jennie M Scarvell

Final approval of the version: M Denika C Silva, Diana M Perriman, Angela M Fearon, Jacqui M Couldrick, Jennie M Scarvell

Role of the funding source

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Conflict of interest

All the authors declare that there are no conflicts of interest in this work.

Ethics approval

This study is a systematic review and does not involve human participants. Therefore, ethics approval is not applicable.

Data Sharing

No additional data are available.

REFERENCES

1. Harris JD, Brand JC, Cote MP, Faucett SC, Dhawan A. Research Pearls: The significance of statistics and perils of pooling. part 1: clinical versus statistical significance. Arthroscopy 2017; 33: 1102-1112.
2. Terwee CB, Roorda LD, Dekker J, Bierma-Zeinstra SM, Peat G, Jordan KP, et al. Mind the MIC: large variation among populations and methods. J Clin Epidemiol 2010; 63: 524-534.
3. de Vet HC, Terwee CB, Ostelo RW, Beckerman H, Knol DL, Bouter LM. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. Health Qual Life Outcomes 2006; 4: 54-54.
4. de Vet HC, Terwee CB. The minimal detectable change should not replace the minimal important difference. J Clin Epidemiol 2010; 63: 804-805; author reply 806.
5. de Vet HCW, Beckerman H, Terwee CB, Terluin B, Bouter LM. Definition of clinical differences. J Rheumatol 2006; 33: 434.
6. Katz NP, Paillard FC, Ekman E. Determining the clinical importance of treatment benefits for interventions for painful orthopedic conditions. J Orthop Surg Res 2015; 10: 24.
7. Wright A, Hannon J, Hegedus EJ, Kavchak AE. Clinimetrics corner: a closer look at the minimal clinically important difference (MCID). J Man Manip Ther 2012; 20: 160-166.
8. Jaeschke R, Singer J, Guyatt GH. Measurement of health status: Ascertaining the minimal clinically important difference. Control Clin Trials 1989; 10: 407-415.

9. Wells G, Anderson J, Beaton D, Bellamy N, Boers M, Bombardier C, et al. Minimal clinically important difference module: summary, recommendations, and research agenda. J Rheumatol 2001; 28: 452-454.

10. Beard DJ, Harris K, Dawson J, Doll H, Murray DW, Carr AJ, et al. Meaningful changes for the Oxford hip and knee scores after joint replacement surgery. J Clin Epidemiol 2015; 68: 73-79.

11. King MT, Dueck AC, Revicki DA. Can methods developed for interpreting group-level patient-reported outcome data be applied to individual patient management? Med Care 2019; 57.

12. Clement ND, Bardgett M, Weir D, Holland J, Gerrand C, Deehan DJ. What is the minimum clinically important difference for the WOMAC index after TKA? Clin Orthop Relat Res 2018; 476: 2005-2014.

13. King MT. A point of minimal important difference (MID): a critique of terminology and methods. Expert Rev Pharmacoecon Outcomes Res 2014; 11: 171-184.

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

15. Mills KA, Naylor JM, Eyles JP, Roos EM, Hunter DJ. Examining the minimal important difference of patient-reported outcome measures for individuals with knee osteoarthritis: A model using the Knee Injury and Osteoarthritis Outcome Score. J Rheumatol 2016; 43: 395-404.

16. Make B. How can we assess outcomes of clinical trials: The MCID approach. COPD: J Chronic Obstr Pulm Dis 2007; 4: 191-194.

17. Copay AG, Subach BR, Glassman SD, Polly DW, Jr., Schuler TC. Understanding the minimum clinically important difference: a review of concepts and methods. Spine J 2007; 7: 541-546.

18. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Med Care 2003; 41: 582-592.

19. Wyrwich KW. Minimal important difference thresholds and the standard error of measurement: is there a connection? J. Biopharm. Stat. 2004; 14: 97-110.

20. Martin Bland J, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 327: 307-310.

21. Angst F, Aeschlimann A, Angst J. The minimal clinically important difference raised the significance of outcome effects above the statistical level, with methodological implications for future studies. J Clin Epidemiol 2017; 82: 128-136.

22. Cross M, Smith E, Hoy D, Nolte S, Ackerman I, Fransen M, et al. The global burden of hip and knee osteoarthritis: estimates from the global burden of disease 2010 study. Ann Rheum Dis 2014; 73: 1323.

23. Bellamy N, Kirwan J, Boers M, Brooks P, Strand V, Tugwell P, et al. Recommendations for a core set of outcome measures for future phase III clinical trials in knee, hip, and hand osteoarthritis. Consensus development at OMERACT III. J Rheumatol 1997; 24: 799-802.

24. Moher D, Liberati A, Tetzlaff J, Altman DG, Group atP. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. Ann Intern Med 2009; 151: 264-269.

25. Dawson J, Boller I, Doll H, et al. Minimally important change was estimated for the Manchester-Oxford Foot Questionnaire after foot/ankle surgery. J Clin Epidemiol 2014;67(6):697-705.

26. Devji T, Carrasco-Labra A, Qasim A, Phillips M, Johnston BC, Devasenapathy N, et al. Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study. BMJ 2020; 369: m1714.

27. Hao Q, Devji T, Zeraatkar D, Wang Y, Qasim A, Siemieniuk RAC, et al. Minimal important differences for improvement in shoulder condition patient-reported outcomes: a systematic review to inform a BMJ Rapid Recommendation. BMJ Open 2019; 9: e028777.

28. Programme CAS. CASP (Qualitative) Checklist. [online] 2018.

29. Long HA, French DP, Brooks JM. Optimising the value of the critical appraisal skills programme (CASP) tool for quality appraisal in qualitative evidence synthesis. Research Methods in Medicine & Health Sciences 2020; 1: 31-42.

30. NHLBI N-NH, Lung, and Blood Institute, National Institute of Health quality assessment tool for before-after (pre-post) studies with no control group. 2013.

31. Haby MM, Chapman E, Clark R, Barreto J, Reveiz L, Lavis JN. What are the best methodologies for rapid reviews of the research evidence for evidence-informed decision making in health policy and practice: a rapid review. Health Res Policy Syst 2016; 14: 83.

32. McCarthy CJ, Oldham JA. The reliability, validity and responsiveness of an aggregated locomotor function (ALF) score in patients with osteoarthritis of the knee. Rheumatology (Oxford) 2004; 43: 514-517.

33. Higgins JPT CJ, Cumpston M, Li T, Page MJ, Welch VA. Cochrane handbook for systematic reviews of interventions. version 6.0 edition2019.

34. 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; 3: e003365.

35. Klokker L, Christensen R, Waehrens EE, Bandak E, Bartholdy C, Bliddal H, et al. Dynamic weight-bearing assessment of pain in knee osteoarthritis: construct validity, responsiveness, and interpretability in a research setting. Health Qual. Life Outcomes 2016; 14: 91.

36. Angst F, Benz T, Lehmann S, Aeschlimann A, Angst J. Multidimensional minimal clinically important differences in knee osteoarthritis after comprehensive rehabilitation: a prospective evaluation from the Bad Zurzach Osteoarthritis Study. RMD Open 2018; 4: e000685.

37. Hmamouchi I, Allali F, Tahiri L, et al. Clinically important improvement in the WOMAC and predictor factors for response to non-specific non-steroidal anti-inflammatory drugs in osteoarthritic patients: a prospective study. BMC Res Notes 2012;5:58

38. Mostafaee N, Nourollahi F, Mostamand J, Negahban H. Responsiveness and the minimal important change of Knee injury and Osteoarthritis Outcome Score in Persian patients with knee osteoarthritis following physiotherapy intervention. Physiother Theory Pract 2021: 1-10.

39. Ornetti P, Dougados M, Paternotte S, Logeart I, Gossec L. Validation of a numerical rating scale to assess functional impairment in hip and knee

osteoarthritis: comparison with the WOMAC function scale. Ann Rheum Dis 2011; 70: 740-746.

40. 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; 41: 509-515.

41. Williams VJ, Piva SR, Irrgang JJ, Crossley C, Fitzgerald GK. Comparison of reliability and responsiveness of patient-reported clinical outcome measures in knee osteoarthritis rehabilitation. J Orthop Sports Phys Ther 2012; 42: 716-723.

42. Perrot S, Bertin P. "Feeling better" or "feeling well" in usual care of hip and knee osteoarthritis pain: determination of cutoff points for patient acceptable symptom state (PASS) and minimal clinically important improvement (MCII) at rest and on movement in a national multicenter cohort study of 2414 patients with painful osteoarthritis. J Pain 2013; 154: 248-256.

43. Lee AC, Driban JB, Price LL, Harvey WF, Rodday AM, Wang C. responsiveness and minimally important differences for 4 Patient-Reported Outcomes Measurement Information System short forms: physical function, pain interference, depression, and anxiety in knee osteoarthritis. J Pain 2017; 18: 1096-1110.

44. Salottolo K, Stahl E. Minimal clinically important improvement response in patients with severe osteoarthritis of the knee: Short report from a survey of clinicians. J Orthop 2018; 15: 424-425.

45. Alghadir A, Anwer S, Brismée J-M. The reliability and minimal detectable change of Timed Up and Go test in individuals with grade 1-3 knee osteoarthritis. BMC Musculoskelet Disord 2015; 16: 174-174.

46. Alghadir A, Anwer S, Iqbal ZA, Alsanawi HA. Cross-cultural adaptation, reliability and validity of the Arabic version of the reduced Western Ontario and McMaster Universities Osteoarthritis index in patients with knee osteoarthritis. Disabil Rehabil 2016; 38: 689-694.

47. Alghadir AH, Anwer S, Iqbal ZA. The psychometric properties of an Arabic numeric pain rating scale for measuring osteoarthritis knee pain. Disabil Rehabil 2016; 38: 2392-2397.

48. Alghadir AH, Al-Eisa ES, Anwer S. Cross-cultural adaptation and psychometric analysis of the Arabic version of the oxford knee score in adult male with knee osteoarthritis. BMC Musculoskelet Disord 2017; 18: 190.

49. Alghadir AH, Anwer S, Iqbal A, Iqbal ZA. Test-retest reliability, validity, and minimum detectable change of visual analog, numerical rating, and verbal rating scales for measurement of osteoarthritic knee pain. J Pain Res 2018; 11: 851-856.

50. Suwit A, Rungtiwa K, Nipaporn T. Reliability and validity of the Osteoarthritis Research Society International minimal core set of recommended performance-based tests of physical function in knee osteoarthritis in community-dwelling adults. Malays J Med Sci 2020; 27: 77-89.

51. Baert IAC, Lluch E, Struyf T, Peeters G, Van Oosterwijck S, Tuynman J, et al. Inter- and intrarater reliability of two proprioception tests using clinical applicable measurement tools in subjects with and without knee osteoarthritis. Musculoskelet Sci Pract 2018; 35: 105-109.

52. Brisson NM, Stratford PW, Maly MR. Relative and absolute test-retest reliabilities of biomechanical risk factors for knee osteoarthritis progression: benchmarks for meaningful change. Osteoarthr Cartil 2018; 26: 220-226.

53. Callaghan MJ, McCarthy CJ, Oldham JA. The reliability of surface electromyography to assess quadriceps fatigue during multi joint tasks in healthy and painful knees. J Electromyogr Kinesiol 2009; 19: 172-180.

54. Suhail A, Chaudhary S. Test-retest reliability and minimum detectable change of 2-minute walk test among individuals with knee osteoarthritis. J Clin Diagnostic Res 2021; 15: 4-7.

55. Hoglund LT, Folkins E, Pontiggia L, Knapp MW. The validity, reliability, measurement error, and minimum detectable change of the 30-second fast-paced walk test in persons with knee osteoarthritis: a novel test of short-distance walking ability. ACR Open Rheumatol 2019; 1: 279-286.

56. Iijima H, Shimoura K, Eguchi R, Aoyama T, Takahashi M. Concurrent validity and measurement error of stair climb test in people with pre-radiographic to mild knee osteoarthritis. Gait Posture 2019; 68: 335-339.

57. Jansen MP, Welsing PMJ, Vincken KL, Mastbergen SC. Performance of knee image digital analysis of radiographs of patients with end-stage knee osteoarthritis. Osteoarthr Cartil 2021; 29: 1530-1539.

58. Kean CO, Birmingham TB, Garland SJ, Bryant DM, Giffin JR. Minimal detectable change in quadriceps strength and voluntary muscle activation in patients with knee osteoarthritis. Arch Phys Med Rehabil 2010; 91: 1447-1451.

59. Klokker L, Christensen R, Osborne R, Ginnerup E, Waehrens EE, Bliddal H, et al. Dynamic weight-bearing assessment of pain in knee osteoarthritis: a reliability and agreement study. Qual Life Res 2015; 24: 2985-2992.

60. McCarthy CJ, Callaghan MJ, Oldham JA. The reliability of isometric strength and fatigue measures in patients with knee osteoarthritis. Man Ther 2008; 13: 159-164.

61. Monticone M, Sconza C, Portoghese I, Nishigami T, Wand BM, Sorrentino G, et al. Cross-cultural adaptation, reliability and validity of the Fremantle Knee Awareness Questionnaire in Italian subjects with painful knee osteoarthritis. Health Qual Life Outcomes 2021; 19: 114.

62. Nalbant A, Unver B, Karatosun V. Test–retest reliability of the L-test in patients with advanced knee osteoarthritis. Physiother Theory Pract 2021: 1-5.

63. Naylor J M, Hayen A, Davidson E, Hackett D, Harris I A, Kamalasena G, et al. Minimal detectable change for mobility and patient-reported tools in people with osteoarthritis awaiting arthroplasty. BMC Musculoskelet Disord 2014: 15, 235.

64. Parveen H, Noohu MM. Evaluation of psychometric properties of Tinetti performance-oriented mobility assessment scale in subjects with knee osteoarthritis. Hong Kong Physiother J 2017; 36: 25-32.

65. Peter WF, de Vet HCW, Terwee CB. Reliability of the Animated Activity Questionnaire for assessing activity limitations of patients with hip and knee osteoarthritis. Musculoskeletal Care 2018; 16: 363-369.

66. Piva SR, Fitzgerald GK, Irrgang JJ, Bouzubar F, Starz TW. Get up and go test in patients with knee osteoarthritis. Arch Phys Med Rehabil. 2004;85(2):284-289.

67. Srimurugan Pratheep N, Madeleine P, Arendt-Nielsen L. Relative and absolute test-retest reliabilities of pressure pain threshold in patients with knee osteoarthritis. Scand J Pain 2018; 18: 229-236.

68. Takacs J, Carpenter MG, Garland SJ, Hunt MA. Test re-test reliability of centre of pressure measures during standing balance in individuals with knee osteoarthritis. Gait Posture 2014; 40: 270-273.

69. Tevald MA, Murray A, Luc BA, Lai K, Sohn D, Pietrosimone B. Hip abductor strength in people with knee osteoarthritis: A cross-sectional study of reliability and association with function. Knee 2016; 23: 57-62.

70. Tse CTF, Charlton JM, Lam J, Ho J, Bears J, Serek A, et al. Smartphone inclinometry is a valid and reliable tool for measuring frontal plane tibial alignment in healthy and osteoarthritic knees. Phys Ther 2021; 101: pzab091.

71. Turcot K, Aissaoui R, Boivin K, Hagemeister N, Pelletier M, de Guise JA. Test-retest reliability and minimal clinical change determination for 3-dimensional tibial and femoral accelerations during treadmill walking in knee osteoarthritis patients. Arch Phys Med Rehabil 2008; 89: 732-737.

72. van der Straaten R, Wesseling M, Jonkers I, Vanwanseele B, Bruijnes A, Malcorps J, et al. Discriminant validity of 3D joint kinematics and centre of mass displacement measured by inertial sensor technology during the unipodal stance task. PLoS One 2020; 15: e0232513.

73. Yuruk Z O, Aytar A, Tüzün E H, Eker L, Yüksel I, Morton N A. Acceptability, reliability and validity of the turkish version of the De morton mobility index in elderly patients with knee osteoarthritis. Turk Geriatri Dergisi 2014; 17.

74. Ravaud P, Giraudeau B, Auleley GR, Edouard-Noël R, Dougados M, Chastang C. Assessing smallest detectable change over time in continuous structural outcome measures: application to radiological change in knee osteoarthritis. J Clin Epidemiol 1999; 52: 1225-1230.

75. Kanko LE, Birmingham TB, Bryant DM, Gillanders K, Lemmon K, Chan R, et al. The star excursion balance test is a reliable and valid outcome measure for patients with knee osteoarthritis. Osteoarthritis Cartilage 2019; 27: 580-585.

76. Mutlu EK, Ozdincler AR. Reliability and responsiveness of algometry for measuring pressure pain threshold in patients with knee osteoarthritis. J Phys Ther Sci. 2015;27(6):1961-1965.

77. Hunter DJ, Conaghan PG, Peterfy CG, Bloch D, Guermazi A, Woodworth T, et al. Responsiveness, effect size, and smallest detectable difference of Magnetic Resonance Imaging in knee osteoarthritis. Osteoarthritis Cartilage 2006; 14 Suppl A: A112-115.

78. Motyl JM, Driban JB, McAdams E, Price LL, McAlindon TE. Test-retest reliability and sensitivity of the 20-meter walk test among patients with knee osteoarthritis. BMC Musculoskelet Disord 2013; 14: 166

79. Perkins NJ, Schisterman EF. The Inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. Am J Epidemiol 2006; 163: 670-675.

80. Redelmeier DA, Lorig K. Assessing the clinical importance of symptomatic improvements. An illustration in rheumatology. Arch Intern Med 1993; 153: 1337-1342.

81. Celik D, Coban O, Kilicoglu O. Minimal clinically important difference of commonly used hip-, knee-, foot-, and ankle-specific questionnaires: a systematic review. J Clin Epidemiol 2019; 113: 44-57.

82. Maredupaka S, Meshram P, Chatte M, Kim WH, Kim TK. Minimal clinically important difference of commonly used patient-reported outcome measures in total knee arthroplasty: review of terminologies, methods and proposed values. Knee Surg Relat Res 2020; 32: 19.

83. MacKay C, Clements N, Wong R, Davis AM. A systematic review of estimates of the minimal clinically important difference and patient acceptable symptom

state of the Western Ontario and McMaster Universities Osteoarthritis Index in patients who underwent total hip and total knee replacement. Osteoarthritis Cartilage 2019; 27: 1408-1419.

84. Escobar A, Riddle DL. Concordance between important change and acceptable symptom state following knee arthroplasty: the role of baseline scores. Osteoarthritis Cartilage 2014; 22: 1107-1110.

85. Angst F, Aeschlimann A, Steiner W, Stucki G. Responsiveness of the WOMAC osteoarthritis index as compared with the SF-36 in patients with osteoarthritis of the legs undergoing a comprehensive rehabilitation intervention. Ann Rheum Dis 2001; 60: 834-840.

86. Bellamy N. Pain assessment in osteoarthritis: experience with the WOMAC osteoarthritis index. Semin. Arthritis Rheum 1989; 18: 14-17.

87. Schwartz CE, Andresen EM, Nosek MA, Krahn GL. Response shift theory: important implications for measuring quality of life in people with disability. Arch Phys Med Rehabil 2007; 88: 529-536.

88. Guyatt GH, Norman GR, Juniper EF, Griffith LE. A critical look at transition ratings. J Clin Epidemiol 2002; 55: 900-908.

89. Barrett B, Brown D, Mundt M, Brown R. Sufficiently important difference: expanding the framework of clinical significance. Med Decis Making 2005; 25: 250-261.

90. Ferreira M. Research Note: The smallest worthwhile effect of a health intervention. J Physiother 2018; 64: 272-274.

91. Ferreira ML, Herbert RD, Ferreira PH, Latimer J, Ostelo RW, Nascimento DP, et al. A critical review of methods used to determine the smallest worthwhile effect of interventions for low back pain. J Clin Epidemiol 2012; 65: 253-261.

92. Wofford N. Smallest worthwhile effect values for pain and function after a total knee replacement. health related sciences, vol. PhD: Virginia Commonwealth University 2020.

93. Bellamy N, Hochberg M, Tubach F, Martin-Mola E, Awada H, Bombardier C, et al. Development of multinational definitions of minimal clinically important improvement and patient acceptable symptomatic state in osteoarthritis. Arthritis Care Res (Hoboken) 2015; 67: 972-980.

94. Kvien TK, Heiberg T, Hagen KB. Minimal clinically important improvement/difference (MCII/MCID) and patient acceptable symptom state (PASS): what do these concepts mean?. Ann Rheum Dis. 2007;66 Suppl 3(Suppl 3):iii40-iii41. doi:10.1136/ard.2007.079798

95. Tubach F, Ravaud P, Baron G, Falissard B, Logeart I, Bellamy N, et al. Evaluation of clinically relevant states in patient reported outcomes in knee and hip osteoarthritis: the patient acceptable symptom state. Ann Rheum Dis 2005; 64: 34-37.

96. Turner D, Schünemann HJ, Griffith LE, Beaton DE, Griffiths AM, Critch JN, et al. Using the entire cohort in the receiver operating characteristic analysis maximizes precision of the minimal important difference. J Clin Epidemiol 2009; 62: 374-379.

97. Lurie F, Kistner RL. In prospective study using Specific Quality of Life & Outcomes Response-Venous (SQOR-V) questionnaire the recall bias had the same magnitude as the minimally important difference. Qual Life Res 2011; 20: 1589-1593.

List of tables

Figure legends

Figure 1: Flow diagram of study selection (PRISMA, Preferred Reporting Items for Systematic reviews and Meta-Analyses)[24]

List of Supplements

Supplement 1: Quality of anchor studies assessed using the Devji (2020) credibility instrument[26]

Supplement 2: Quality of consensus studies assessed using the Critical Appraisal Screening Program (CASP)-qualitative checklist[28]

Supplement 3: Quality of distribution studies assessed using the National Institute of Health (NIH) quality assessment tool for before-after (pre-post) studies with no control group[30]

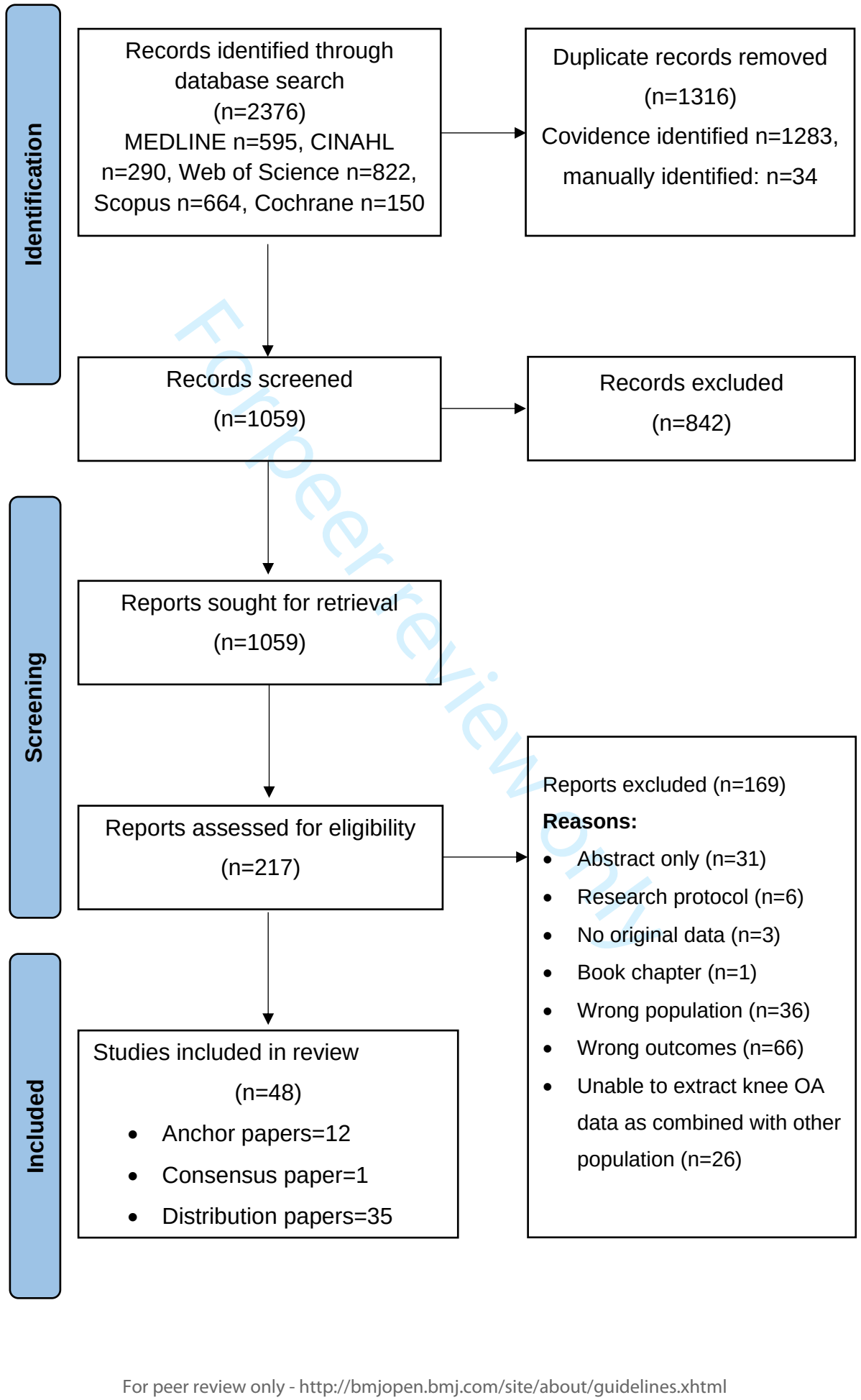
Supplement 4: Anchor properties of included anchor studies

Supplement 5: The effect of follow-up time on **A.** Minimal Clinically Important Change (MCIC) and **B.** Minimal Clinically Important Difference (MCID)

†: multiple MCIC estimates using two different anchor questions; *: multiple MCID estimates using two different anchor questions and different calculation methods

KOOS: Knee injury and Osteoarthritis Outcome Score, QOL: Quality of Life, KOS: Knee Outcome Survey, ADL: Activities of Daily Living, LEFS: Lower Extremity Functional Scale, WOMAC: Western Ontario and McMaster Universities Arthritis Index

Supplement 6: Minimum Detectable Change (MDC) values of knee osteoarthritis outcome tools derived using the distribution method



Supplement 1: Quality of anchor studies assessed using the Devji (2020) credibility instrument[26]

Study	Item 1*	Item 2#	*Item 3#*	Item 4#	Item 5#	The overall quality of the studies**
Angst, 2018[36]	Yes	Impossible to tell	Definitely No	To a greater extent	To a great extent	**High**
Harris, 2013[34]	Yes	To a great extent	Impossible to tell	Not so much	Definitely No	**Low**
Hmamouchi, 2012[37]	Yes	Not so much	Impossible to tell	To a greater extent	a greater extent	**High**
Klokker, 2016[35]	Yes	To a great extent	Impossible to tell	Definitely No	Not so much	**Low**
Lee, 2017[43]	Yes	Definitely Yes	Impossible to tell	To a greater extent	mpossible to tell	**High**
Mills, 2016[15]	Yes	Definitely Yes	Not so much	To a greater extent	Definitely Yes	**High**
Mostafaee, 2021[38]	Yes	To a great extent	To a great extent	To a great extent	Not so much	**High**
Ornetti, 2011[39]	Yes	To a great extent	Impossible to tell	To a greater extent	Definitely No	**High**
Perrot, 2013[42]	Yes	To a great extent	Not so much	Definitely Yes	Not so much	**High**
Singh, 2014[40]	Yes	To a great extent	Impossible to tell	Not so much	o a great extent	**High**
Tubach, 2005[14]	Yes	To a great extent	Impossible to tell	Definitely Yes	Not so much	**High**
Williams, 2012[41]	Yes	Not so much	Impossible to tell	To a greater extent	Definitely Yes	**High**

Item 1: Is the patient or necessary proxy responding directly to both the PROM and the anchor?
Item 2: Is the anchor easily understandable and relevant for patients or a necessary proxy?
Item 3: Has the anchor shown a good correlation with the PROM?
Item 4: Is the MID precise?
Item 5: Does the threshold or difference between groups on the anchor used to estimate the MID reflect a small but important difference?

* The responses to items: Yes, No, Impossible to tell
# The responses to items: Definitely yes, To a great extent, Not so much, Definitely no, Impossible to tell
** overall quality: three of the five criteria were met “Yes” or “definitely yes” or “to a great extent”, the paper is of “high” quality If not, hat the paper is of “low” quality[27]

Low quality (credibility) studies are shaded.

PROM: Patient-Reported Outcome Measure, MID-Minimal Important Difference (In this study minimal clinically important change/difference)

Supplement 2: Quality of consensus studies assessed using the Critical Appraisal Screening Program (CASP)-qualitative checklist.

Study (Author, Year)	Item 1 #	Item 2 #	Item 3 #	Item 4 #	Item 5 #	Item 6 #	Item 7 #	Item 8 #	Item 9 #	Item 10	The overall quality of the studies µ
Salottolo, 2018[44]	Yes	Yes	Can't tell	Can't tell	Yes	Yes	Can't tell	Yes	Yes	Researcher discusses the the contribution the study makes to existing knowledge or understanding	**Moderate**

Item 1: Was there a clear statement of the aims of the research?
Item 2: Is a qualitative methodology appropriate?
Item 3: Was the research design appropriate to address the aims of the research?
Item 4: Was the recruitment strategy appropriate to the aims of the research?
Item 5: Was the data collected in a way that addressed the research issue?
Item 6: Has the relationship between researcher and participants been adequately considered?
Item 7: Have ethical issues been taken into consideration?
Item 8: Was the data analysis sufficiently rigorous?
Item 9: Is there a clear statement of findings?
Item 10: How valuable is the research?

The responses to items: Yes, Can't Tell, No
µ Overall quality of study: high", "moderate" or "low" was evaluated by the review team based on how reliable and credible each study was, without specific rules

Supplement 3: Quality of distribution studies assessed using the National Institute of Health (NIH) quality assessment tool for before-after (pre-post) studies with no control group[30]

Study (Author, Year)	Criteria 1 #	Criteria 2 #	Criteria 3 #	Criteria 4 #	Criteria 5 #	Criteria 6 #	Criteria 7 #	Criteria 8 #	Criteria 9 #	Criteria 10 #	Criteria 11 #	Criteria 12 #	Overall quality µ
Alghadir, 2017[50]	Yes	Yes	Yes	Yes	Other	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Alghadir, 2018[51]	Yes	Yes	Yes	Yes	Other	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Alghadir, 2016 b	Yes	No	Yes	Yes	Yes	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Alghadir, 2017	Yes	Yes	Yes	Yes	Other	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Baert, 2018[53]	Yes	No	Yes	Yes	Other	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Baert, 2018	Yes	Yes	Yes	Yes	Other	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Brisson, 2018[54]	Yes	Yes	Yes	Yes	Yes	Other	Yes	Other	Yes	Yes	Yes	Other	**Good**
Callaghan, 2009[55]	Yes	No	Yes	Yes	Yes	Other	Yes	Other	Other	Yes	Yes	Other	**Fair**
Hoglund, 2019[57]	Yes	Yes	Yes	Yes	Yes	Other	Yes	Other	Yes	Yes	Yes	Other	**Good**
Hunter, 2006[79]	Yes	Yes	Yes	Yes	Other	Other	Yes	Other	Other	Yes	Yes	Other	**Fair**
Ijima, 2019[58]	Yes	Yes	Yes	Yes	Other	Yes	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Jansen, 2021[59]	Yes	Yes	Yes	Yes	Other	Yes	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Kanko, 2019[77]	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Other	Yes	Yes	Yes	Other	**Good**
Kean, 2010[60]	Yes	Yes	Yes	Yes	Other	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Klokker, 2015[61]	Yes	Yes	Yes	Yes	Yes	Other	Yes	Other	Yes	Yes	Yes	Other	**Good**
McCarthy, 2004[32]	Yes	Yes	Yes	Yes	Other	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
McCarthy, 2008[62]	Yes	Yes	No	No	Other	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Monticone, 2021[63]	Yes	Yes	Yes	Yes	Yes	Other	Yes	Other	Yes	Yes	Yes	Other	**Good**
Motyl, 2013[80]	Yes	Yes	Yes	Yes	Yes	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Mutlu, 2015[78]	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Other	Yes	Yes	Yes	Other	**Good**
Nalbant, 2021[64]	Yes	Yes	Yes	Yes	Yes	Other	Yes	Other	Yes	Yes	Yes	Other	**Good**
Naylor, 2014[65]	Yes	Yes	Yes	Yes	Other	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Parveen, 2017[66]	Yes	Yes	Yes	Yes	Yes	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Peter, 2018[67]	Yes	Yes	Yes	Yes	Other	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Piva, 2004[68]	Yes	Yes	Yes	Yes	Other	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Pratheep, 2018[69]	Yes	Yes	Yes	Yes	Other	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Ravaud, 1999[76]	Yes	Yes	Yes	Yes	Other	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**

Study (Author, Year)	Criteria 1 #	Criteria 2 #	Criteria 3 #	Criteria 4 #	Criteria 5 #	Criteria 6 #	Criteria 7 #	Criteria 8 #	Criteria 9 #	Criteria 10 #	Criteria 11 #	Criteria 12 #	Overall quality µ
Suhail and Chaudhary, 2021[56]	Yes	Yes	Yes	Yes	Yes	Other	Yes	Other	Yes	Yes	Yes	Other	**Good**
Suwit, 2020[52]	Yes	Yes	Yes	Yes	Other	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Takacs, 2014	Yes	Yes	Yes	Yes	Other	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Tevald, 2016	Yes	Yes	Yes	Yes	Other	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Takacs, 2014[70]	Yes	Yes	Yes	Yes	Yes	Other	Yes	Other	Yes	Yes	Yes	Other	**Good**
Tevald, 2016[71]	Yes	No	Yes	Yes	Other	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Tse, 2021[72]	Yes	Yes	Yes	Yes	Other	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Turcot, 2008[73]	Yes	Yes	Yes	Yes	Yes	Other	Yes	Other	Yes	Yes	Yes	Other	**Good**

The description of the criteria is given below.
Criteria 1: Was the study question or objective clearly stated?
Criteria 2: Were eligibility/selection criteria for the study population prespecified and clearly described?
Criteria 3: Were the participants in the study representative of those who would be eligible for the test/service/intervention in the general or clinical population of interest?
Criteria 4: Were all eligible participants that met the prespecified entry criteria enrolled?
Criteria 5: Was the sample size sufficiently large to provide confidence in the findings?
Criteria 6: Was the test/service/intervention clearly described and delivered consistently across the study population?
Criteria 7: Were the outcome measures prespecified, clearly defined, valid, reliable, and assessed consistently across all study participants?
Criteria 8: Were the people assessing the outcomes blinded to the participants' exposures/interventions?
Criteria 9: Was the loss to follow-up after baseline 20% or less? Were those lost to follow-up accounted for in the analysis?
Criteria 10: Did the statistical methods examine changes in outcome measures from before to after the intervention? Were statistical tests done that provided p values for the pre-to-post changes?
Criteria 11: Were outcome measures of interest taken multiple times before the intervention and multiple times after the intervention (i.e., did they use an interrupted time-series design)?
Criteria 12: If the intervention was conducted at a group level (e.g., a whole hospital, a community, etc.) did the statistical analysis take into account the use of individual-level data to determine effects at the group level?

The responses to items: Yes, No, Other (CD, NR, NA)*, *CD, cannot determine; NA, not applicable; NR, not reported
µ Overall quality: Good, Fair, or Poor was evaluated by the review team based on how reliable and credible each study was, without specific rules as recommended by the tool

Supplement 4: Anchor properties of included anchor studies

Study	Description of anchor	Anchor question	Anchor properties/responses	Definitions of MCIC/MCID using transition question	Reported terminology in the study
Angst, 2018[36]	GRC transition (5 points)	NR	much worse, slightly worse, almost equal, slightly better, much better	Difference between the "slightly better" group and the "almost equal" group= MCID	Minimal Clinically Important Difference
Harris, 2013[34]	GRC responses (3 points)	Compared to one week before your clinical visit, please indicate how much your knee problem has changed?	1. My knee has got better, 2. My knee has stayed the same, 3. My knee has got worse	Mean change in the group "my knee has got better"= MCIC and the difference in the change score between groups responded with " my knee has stayed the same" and "my knee has got better"= MCID	Minimal Important Change, Minimal Important Difference
Hmamouchi, 2012[37]	GRC transition (5 points)	How do you feel in general today as compared to six weeks earlier as far as your osteoarthritis is compared?	much better, slightly better, no change, slightly worse, much worse	Difference between the mean effects of "slightly better' group and "no change" groups= MCID	Minimal Clinically Important Difference for improvement
Klokker, 2016[35]	modified GRC (3 responses and a GRC spanning from -7 to +7)	Did your knee pain change since you entered this project?	unchanged, better, worse: if it is better or worse, bring up a scale spanning from -7 to +7 (-7 (worst) to +7 (best) scale)	Difference of a score of at least 2 (+2: little better: -2: little worse) and no change (score of 0, +1 (almost the same or hardly any better), -1 (worse at all))= MCID	Minimal Important Change
Lee, 2018[43]	Multiple anchors: 36 item Short Form subscales- physical functioning, social role functioning, energy and vitality, bodily pain, mental health, physical role functioning, emotional role functioning, general	NR	NR	Prospective change for people achieving previously established MCID on legacy comparators: MCID in a single item anchor e.g. Patient Global Assessment was defined as range= MCID to 1+MCID, MCID in a multi-teem anchor e.g. SF36 range= MCID to 2XMCID	Minimally Important Difference

Study	Description of anchor	Anchor question	Anchor properties/responses	Definitions of MCIC/MCID using transition question	Reported terminology in the study
	health perception (0-100 scale, 0: best, 100: worse); WOMAC-pain and function (0-100 scales,0: best, 100: worse); Back depression (0-63, 0: best, 63: worse); Perceived Stress scale (0-40, 0: best, 40: worse); Patient Global Assessment in a 0-10 cm visual analogue scale (higher number= greater perception of disease activity); Six-minute walk test (in meters); 20-meter walk test (in seconds)				
Mills, 2016[15]	Global transition GRC-7-point Likert scale	Two anchor question 1. Compared with when I started this program, my walking on level ground has (walking anchor) and 2. Compared with when I started this program, my knee had (knee health anchor):	much improved, moderately improved, slightly improved, not changed, slightly worse, moderately worse, much worse	Mean change within slightly improved group = MCIC And Difference between participants who responded slightly, moderately, much improved as the "improved group" and no change or worse "non-improved group" =MCID	Minimal Important Difference-Mean change (Jaeschke) method Minimal Important Difference-Mean change (Redelmeier and Lorig) method, ROC method (Youden method and 80% specific method)
Mostafee, 2021[38]	GRC- 7points Likert scale	How did your knee status change compared to the	1. very much worse; 2. much worse; 3. slightly worse; 4. no change; 5.	The difference between improved group (6 and 7) and not improved group (1,2,3,4 and 5)=MCID	Minimal Important Change

Study	Description of anchor	Anchor question	Anchor properties/responses	Definitions of MCIC/MCID using transition question	Reported terminology in the study
		beginning of the physiotherapy intervention?	slightly better; 6. much better; and 7. very much better		
Ornetti, 2011[39]	GRC- (3points Likert scale, then again in a 4-point Likert scale	specific question NR. Two anchor questions. 1. The degree of improvement of global state 2. The degree of improvement of functional state	The degree of improvement of global state (global MCII) on a 3-point Likert scale (worsened function, no change, improved function). Then, among the patients who improved, the degree of improvement was scored on a scale (poor, fair, good, excellent)	75th centile of the absolute change in score among patients who responded the improvement as good or excellent (improved group) = MCIC	Minimal Clinically Important Improvement
Perrot, 2013[42]	GRC (5 points)	Taking into account the pain due to your arthritis in the last 24 hours and the pain you experienced initially, how has your pain changed?	Patients who said that their pain had improved were asked to rate the level of improvement on a 5-point Likert scale extending from very large improvement to no improvement at all	75th percentile of the distribution of the pain intensity difference (day 0 to 7) for patients considering their improvement to be at least moderate= MCIC	Minimal Clinically Important Improvement
Singh, 2014[40]	GRC (5 points)	Since the last time you completed the survey 2 weeks ago, would you say your knee arthritis is:	A great deal better, somewhat better, about the same, somewhat worse, a great deal worse	Mean change between baseline and follow-up of the group who said somewhat better= MCIC	Minimal Clinically Important Difference for improvement

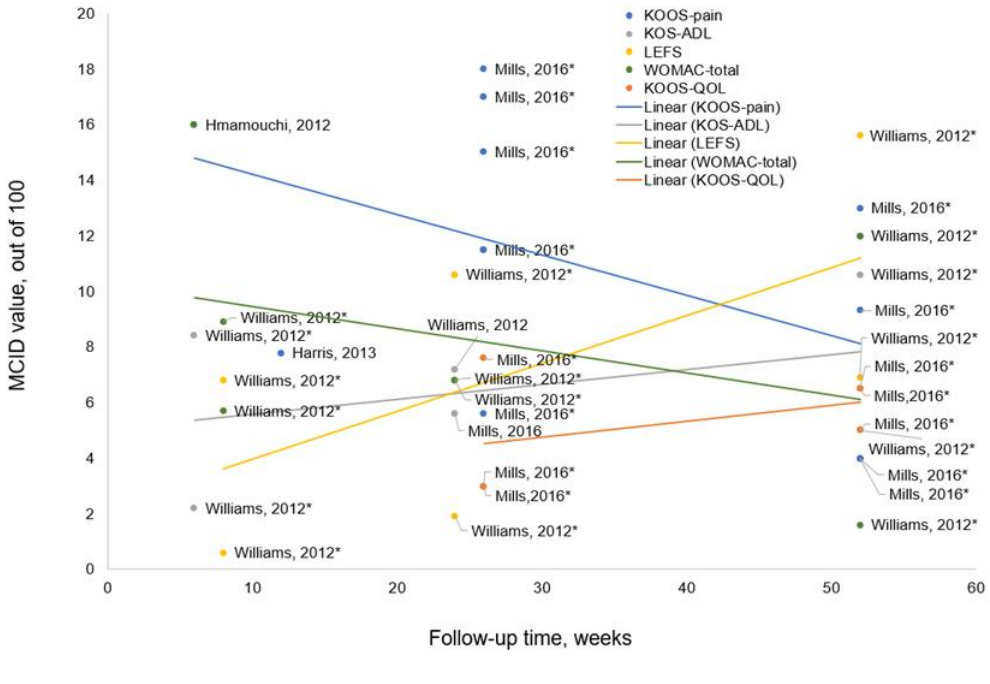
Study	Description of anchor	Anchor question	Anchor properties/responses	Definitions of MCIC/MCID using transition question	Reported terminology in the study
Tubach, 2005[14]	GRC (5 points)	Response to NSAID treatment	None= no good at all, ineffective drug; poor= some effect but unsatisfactory; fair= reasonable effect but could be better; good= satisfactory effect with occasional episodes of pain or stiffness; excellent= ideal response, virtually pain-free	75th centile of the change in score of the group who responded as “good” using logistic regression=MCIC	Minimal Clinically Important Improvement
Williams, 2012[41]	GRC questionnaire on a 15-level scale (15 points)	Not specified To rate the extent to which they perceive their condition as having changed over time	GRC on a 15- level scale. The GRC ranges from 1= a very great deal better to 8=about the same to 15= very great deal worse	Difference between subjects who perceived "improved" (those with a GRC between 1 (very great deal better) and 5 (somewhat better)) from subjects who perceived "not improved" (those with between 6 (a little better) and 15 (a very great deal worse)) = MCID	Minimal Clinically Important Difference

MCIC: Minimal Clinically Important Change, MCID: Minimal Clinically Important Difference, GRC: Global Rating of Change, NR: Not Reported, CI: Confidence Interval, WOMAC: Western Ontario and McMaster Universities Arthritis Index, SF36: Short Form-36, NSAID: Non-Steroid Anti-Inflammatory Drug

Supplement 5: The effect of follow-up time on **A.** Minimal Clinically Important Change (MCIC) and **B.** Minimal Clinically Important Difference (MCID)

A.

B.



†: multiple MCIC estimates using two different anchor questions; *: multiple MCID estimates using two different anchor questions and different calculation methods

KOOS: Knee injury and Osteoarthritis Outcome Score, QOL: Quality of Life, KOS: Knee Outcome Survey, ADL: Activities of Daily Living, LEFS: Lower Extremity Functional Scale, WOMAC: Western Ontario and McMaster Universities Arthritis Index

Supplement 6: Minimum Detectable Change (MDC) values of knee osteoarthritis outcome tools derived using the distribution method

Outcome tool	Score	Median	Minimum	Maximum	Number of estimates	Study
Aggregated locomotor function	N/A	2.3 seconds			1	McCarthy,2004[32]
Animated Activity Questionnaire	100=best	11.2			1	Peter, 2018[65]
BMI using a scale and a stadiometer	N/A	2.6 kg/m^2			1	Brisson, 2018[52]
Bone density-femur mean lateral	N/A	0.5 mm	0.4 mm	0.6 mm	2	Jansen, 2021[57]
Bone density-femur mean medial	N/A	0.6 mm Al eq	0.5 mm Al eq	0.7 mm Al eq	2	Jansen, 2021[57]
Bone density-tibia mean lateral	N/A	2.6 mm Al eq	2.4 mm Al eq	2.8 mm Al eq	2	Jansen, 2021[57]
Bone density-tibia mean medial	N/A	0.8 mm Al eq	0.7 mm Al eq	1 mm Al eq	2	Jansen, 2021[57]
Cartilage volume-lateral tibia	N/A	0.6ml	0.5 ml	0.6 ml	2	Hunter, 2006[77]
Cartilage volume-medial Tibia	N/A	0.6 ml	0.4 ml	0.7 ml	2	Hunter, 2006[77]
Cartilage volume-femur	N/A	1.1 ml	0.8 ml	1.3 ml	2	Hunter, 2006[77]
Cartilage volume-patella	N/A	0.9 ml	0.7 ml	1.1 ml	2	Hunter, 2006[77]
Center of mass mediolateral displacement during unipodal stance using 3D motion analysis	N/A	0.01º	0.01º	0.02 º	3	VandeStraaten, 2020[72]
De Motion Mobility Index	100=best	8.0	7.3	8.7	2	Yuruk, 2014[73]
Eight-meter walk time	N/A	1.4 seconds			1	McCarthy,2004[32]
Eminence lateral (knee image)	mm	2.2 mm	2.2 mm	2.2 mm	2	Jansen, 2021[57]
Eminence medial (knee image)	mm	1.8 mm	1.7 mm	1.8 mm	2	Jansen, 2021[57]
Get up and Go test	N/A	1.4 seconds	1.2 seconds	1.5 seconds	2	Piva, 2004[66]
Fremantle Knee Awareness Questionnaire	100=worst	14.4			1	Monticone,2021[61]
Frontal plane tibial alignment using smartphone inclinometer	N/A	3.7º			1	Tse, 2021[70]
Frontal plane tibial alignment using a manual inclinometer	N/A	3.2º			1	Tse, 2021[70]
Hip abductor strength using a hand-held dynameter	N/A	0.3 Nm/kg	0.3 Nm/kg	0.3 Nm/kg	2	Tevald, 2016[69]
Hip flexion-extension during unipodal stance using 3D motion analysis	N/A	2.7 º	2.2 º	4.6 º	3	VandeStraaten, 2020[72]
ICOAP-constant pain	100=worst	51.7	49.6	53.8	2	Singh, 2014[40]
ICOAP-intermittent pain	100=worst	49.8	48.7	50.8	2	Singh, 2014[40]
ICOAP-pain	100=worst	46.6	23.0	49.6	3	Singh, 2014[40], Harris, 2013[34]
KAM impulse in 3D motion analysis	N/A	4.9 Nm*s			1	Brisson, 2018[52]
KAM impulse in 3D motion analysis	N/A	0.4 %BW*HT*s			1	Brisson, 2018[52]

Outcome tool	Score	Median	Minimum	Maximum	Number of estimates	Study
Knee force sense test using a handheld dynamometer-reposition error 20º	N/A	20.6 N	13.2 N	26.7 N	3	Baert, 2018[51]
Knee force sense test using a handheld dynamometer-reposition error 45º	N/A	38.0 N	27.8 N	43.3 N	3	Baert, 2018[51]
Knee force sense test using a handheld dynamometer-reposition error 70º	N/A	32.2 N	22.5 N	49.5 N	3	Baert, 2018[51]
Knee force sense test using a handheld dynamometer-reposition error all	N/A	32.7 N	30.0 N	33.7 N	3	Baert, 2018[51]
Knee joint angle	degrees	2.1º	1.9º	2.3º	2	Jansen, 2021[57]
Knee joint space width	mm	1 mm	0.8 mm	1.2 mm	2	Jansen, 2021[57]
Knee joint space width-lateral	mm	1.7 mm	1.4 mm	2.0 mm	2	Jansen, 2021[57]
Knee joint space width-medial	mm	0.6 mm	0.5 mm	0.8 mm	2	Jansen, 2021[57]
Knee joint space width-minimum	mm	0.8 mm	0.6 mm	0.9mm	2	Jansen, 2021[57]
KOOS-activities of daily living	100=best	19.1	17.4	20.8	2	Naylor, 2014[63]
KOOS-pain	100=best	18.6	17	20.2	2	Naylor, 2014[63]
KOOS-quality of life	100=best	27.8	22.4	39.0	4	Naylor, 2014[65]; Singh, 2014[42]
KOOS-symptoms	100=best	20.2	2.9	24.1	3	Naylor, 2014[63]
KOOS-PS	100=worst	28.3	16.0	35.5	3	Harris, 2013[34]; Singh, 2014[40]
Knee joint position sense test-analogue inclinometer-reposition error 70º	N/A	8.0 º	4.0 º	8.0 º	3	Baert, 2018[51]
Knee joint position sense test-analogue inclinometer-reposition error 45º	N/A	4.0 º	3.0 º	8.0 º	3	Baert, 2018[51]
Knee joint position sense test-analogue inclinometer-reposition error 20º	N/A	4.0 º	3.0 º	4.0 º	3	Baert, 2018[51]
Knee joint position sense test-analogue inclinometer-reposition error all	N/A	3.0 º	3.0 º	4.0 º	3	Baert, 2018[51]
KOS-activities of daily living	100=best	15.8*	10.5	21.0	6	Williams, 2012[41]
L-test	N/A	5.28 seconds			1	Nalbant, 2021[62]
Load frequency using a triaxial accelerometer	N/A	4.3 steps/day			1	Brisson, 2018[52]
LEFS	100=best	18.3*	14.8	22.6	6	Williams, 2012[41]

Outcome tool	Score	Median	Minimum	Maximum	Number of estimates	Study
NRS- pain	100=worst	16.5	13.3	19.6	2	Alghadir, 2016b[47], Alghadir, 2018[49]
Osteophytes-Femur lateral (using knee image)	N/A	7.8 mm^2	5.4 mm^2	10.3 mm^2	2	Jansen, 2021[57]
Osteophytes-Femur medial (using knee image)	N/A	12.3 mm^2	11.2 mm^2	13.4 mm^2	2	Jansen, 2021[57]
Osteophytes-Tibia lateral (using knee image)	N/A	10.5 mm^2	9.4 mm^2	11.6 mm^2	2	Jansen, 2021[57]
Osteophytes-Tibia medial (using knee image)	N/A	11.6 mm^2	11 mm^2	12.2 mm^2	2	Jansen, 2021[57]
OKS-summary	100=best	6.1	6.0	6.2	1	Alghadir, 2017[48], Harris, 2013[34]
Peak KAM in 3D motion analysis	N/A	0.2 Nm/kg			1	Brisson, 2018[52]
Peak KAM in 3D motion analysis	N/A	1.3 %BW*HT			1	Brisson, 2018[52]
Peak KFM in 3Dmotion analysis	N/A	0.4 Nm/kg			1	Brisson, 2018[52]
Peak KFM in 3D motion analysis	N/A	2.3 %BW*HT			1	Brisson, 2018[52]
Pelvic abduction-adduction during unipodal stance using 3D motion analysis	N/A	3.1 º	2.8 º	3.5 º	3	VandeStraaten, 2020[72]
Per cent of voluntary muscle activation using a dynamometer	N/A	6.6%			1	Kean, 2010[58]
Pressure pain threshold-knee- using an algometer	N/A	131.8 kPa	92.9 kPa	196.3 kPa	10	Pratheep, 2018[67]
Pressure pain threshold-medial heel using a handheld pressure algometer	N/A	1.3 lb			1	Mutlu, 2015[76]
Pressure pain threshold- medial knee using a handheld pressure algometer	N/A	1.2 lb			1	Mutlu, 2015[76]
Quadriceps fatigue using surface electromyography-VMO initial median frequency	N/A	11.0 Hz			1	Callghan, 2009[53]
Quadriceps fatigue using surface electromyography-VL initial median frequency	N/A	10.0 Hz			1	Callghan, 2009[53]
Quadriceps fatigue using surface electromyography-RF initial median frequency	N/A	9.0 Hz			1	Callghan, 2009[53]
Quadriceps fatigue using surface electromyography-VMO final median frequency	N/A	7.4 Hz			1	Callghan, 2009[53]
Quadriceps fatigue using surface electromyography-VL final median frequency	N/A	6.5 Hz			1	Callghan, 2009[53]
Quadriceps fatigue using surface electromyography-RF final median frequency	N/A	10.5 Hz			1	Callghan, 2009[53]

Outcome tool	Score	Median	Minimum	Maximum	Number of estimates	Study
Quadriceps fatigue using surface electromyography-VMO median frequency slope	N/A	2207.0 %/min*			1	Callghan, 2009[53]
Quadriceps fatigue using surface electromyography-VL median frequency slope	N/A	4000.0 %/min*			1	Callghan, 2009[53]
Quadriceps fatigue using surface electromyography-RF median frequency slope	N/A	2390.0 %/min*			1	Callghan, 2009[53]
Quadriceps isokinetic strength using a dynamometer-absolute value	N/A	33.9 Nm			1	Kean, 2010[58]
Quadriceps isometric strength using a dynamometer-absolute value	N/A	25.0 Nm	5.5	37.2	3	McCarthy,2008[60]; Brisson, 2018[52]
Quadriceps isometric strength using a dynamometer-normalised to weight	N/A	0.4 Nm/kg	0.3	0.5	2	Brisson, 2018[52]; Kean, 2010[58]
Quadriceps isometric strength using a dynamometer-Normalised to body size	N/A	1.5 %BW*height			1	Kean, 2010[58]
Quadriceps power using a dynamometer	N/A	151.8 W/2.2 W/kg			1	Brisson, 2018[52]
Rectus femoris fatigue slope	N/A	0.6 %/min			1	McCarthy,2008[60]
Rectus femoris initial median frequency	N/A	5.2 Hz			1	McCarthy,2008[60]
Single-leg standing balance test-mediolateral standard deviation	N/A	0.3			1	Takacs, 2014[68]
Single-leg standing balance test-anteroposterior standard deviation	N/A	0.5			1	Takacs, 2014[68]
Single-leg standing balance test-path length	N/A	20.2 cm			1	Takacs, 2014[68]
Single-leg standing balance test-velocity	N/A	0.3 m/seconds			1	Takacs, 2014[68]
Single-leg standing balance test-area	N/A	23.2 cm^2			1	Takacs, 2014[68]
Six minute walk test	N/A	72.7 m	66.3 m	79.0 m	2	Naylor, 2014[63]
Stair ascent and descent time (seven steps (four of 15cm, three of 20cm))	N/A	2.6 seconds			1	McCarthy,2004[32]
Stopwatch-based 11- Step stair climb test	N/A	0.1 seconds	0.1 seconds	0.1 seconds	2	Ijima, 2019[56]
Star excursion balance test-Raw value	N/A	7.4 cm			1	Kanko, 2019[75]
Star excursion balance test-Normalized (raw value/leg length%)	N/A	8.5 %			1	Kanko, 2019[75]
Timed Up and Go test	N/A	1.1 seconds	1.1 seconds	1.1 seconds	2	Alghadir, 2015[45]

Outcome tool	Score	Median	Minimum	Maximum	Number of estimates	Study
Timed Up and Go test (as a ratio of the original measurement)	N/A	41.06%	37.5%	44.6%	2	Naylor, 2014[63]
Tinetti Performance-Oriented Mobility Assessment scale- balance subscale	NR	0.8			1	Parveen, 2017[64]
Tinetti Performance-Oriented Mobility Assessment scale- gait subscale	NR	0.6			1	Parveen, 2017[64]
Tinetti Performance-Oriented Mobility Assessment scale- total	NR	1.0			1	Parveen, 2017[64]
Transferring time (distance of 2m to a chair and sit down, then, walk back to the start)	N/A	1.7			1	McCarthy,2004[32]
Trunk-abduction-adduction during unipodal stance using 3D motion analysis	N/A	2.9 º	2.6 º	2.9 º	3	VandeStraaten, 2020[72]
2-minute walk test	N/A	5.52m			1	Suhail and Chaudhary, 2021[54]
Vastus lateralis fatigue slope	N/A	0.5 %/min			1	McCarthy,2008[60]
Vastus lateralis initial median frequency	N/A	5.8 Hz			1	McCarthy,2008[60]
Vastus medialis oblique fatigue slope	N/A	0.8 %/min			1	McCarthy,2008[60]
Vastus medialis oblique initial median frequency	N/A	6.7 Hz			1	McCarthy,2008[60]
Verbal Rating Scale for pain		5.8			1	Alghadir, 2018[49]
VAS for pain	100=worst	24.0 cm *	8.0 cm	28.0 cm	3	Alghadir, 2018[49]; Naylor,2014[63]
WOMAC-pain	100=best	3.8	3.4	19.0	3	Angst, 2018[36], Alghadir, 2016 a[46]
WOMAC-function	100=best	3.1	3.1	18.7	3	Angst, 2018[36], Alghadir, 2016 a[46]
WOMAC-stiffness	100=best	5.2	4.8	5.6	2	Angst, 2018[36]
WOMAC-functional standing/walking	100=best	3.8	3.5	4.0	2	Angst, 2018[36]
WOMAC-total	100=worst	18.7	11.7	20.8	6	Williams, 2012[43], Alghadir, 2016 a[48]
SF-36-bodily pain	100=best	3.5	3.3	3.7	2	Angst, 2018[36]
SF-36-general health	100=best	2.1	2.0	2.2	2	Angst, 2018[36]
SF-36-mental health	100=best	3.1	2.9	3.4	2	Angst, 2018[36]
SF-36-physical functioning	100=best	2.4	2.3	2.4	2	Angst, 2018[36]

Outcome tool	Score	Median	Minimum	Maximum	Number of estimates	Study
SF-36-role- physical	100=best	11.2	10.9	11.5	2	Angst, 2018[36]
SF-36-social functioning	100=best	4.1	3.7	4.5	2	Angst, 2018[36]
SF-36-vitality	100=best	5.3	5.1	5.6	2	Angst, 2018[36]
20-meter walk test	N/A	0.9 seconds	0.2 seconds	1.6 seconds	2	Motyl, 2013[78]
30-second fast-paced walk test	N/A	3.2 m	0.8 m	5.7 m	2	Hoglund, 2019[55]
40-meter fast-paced test	N/A	16.3			1	Suwit, 2020[50]
30 seconds chair stand test	N/A	21.2			1	Suwit, 2020[50]
3D linear accelerations of the tibia during comfortable walking	N/A	0.3 g	0.1 g	0.8 g	12	Turcot, 2008[71]
3D linear accelerations of the femur during comfortable walking	N/A	0.3 g	0.1 g	1.0 g	12	Turcot, 2008[71]
3D linear accelerations of the tibia at fast speed	N/A	0.4 g	0.1 g	0.9 g	12	Turcot, 2008[71]
3D linear accelerations of the femur at fast speed	N/A	0.2 g	0.0 g	0.6 g	12	Turcot, 2008[71]
9-step stair climb test	N/A					Suwit, 2020[02]

MDC: Minimum Detectable Change, OA: osteoarthritis, BMI: Body Mass Index, MRI: Magnetic Resonance Imaging, 3D: 3-dimensional, N/A: Not Applicable, ICOAP: Intermittent and constant osteoarthritis pain, KOOS: Knee injury and Osteoarthritis Outcome Score, KOOS-PS: Knee injury and Osteoarthritis Outcome Score Physical Function Short form, KOS: Knee Outcome Survey, LEFS: Lower Extremity Functional Scale, NRS: Numeric Rating Scale, OKS: Oxford Knee Score, KAM: Knee adduction moment, KFM: Knee Flexion Moment, VMO: Vastus Medialis Oblique, VL: Vastus Lateralis, RF: Rectus Femoris, VAS: Visual Analog Scale, WOMAC: Western Ontario and McMaster Universities Arthritis Index, SF-36: 36-item Short Form health survey,

All the estimates are out of 100.

The median estimates are shaded in blue

PRISMA 2020 Checklist

Section and Topic	Item #	Checklist item	Location where item is reported
TITLE			
Title	1	Identify the report as a systematic review.	Page 1
ABSTRACT			
Abstract	2	See the PRISMA 2020 for Abstracts checklist.	Page 2 to 3
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	Page 5 to 6
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	Page 7, line 96 to 99
METHODS			
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	Page 8 to 9, line 131 to 150
Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	Page 7, line 108 to 112
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	Page 7 to 8, line 112 to 123
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	Page 8, line 125 to 131
Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	Page 10, line 173 to 182; Page 10, line 170 to 172
Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	Page 10, line 173 to 182
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	Page 10, line 173 to 176
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	Page 9 to 10, line 152 to 171
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	Page 10, line 184 to 188
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	Page 10, line 184 to 188
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	Page 10 to 11, line 184 to 196
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	Page 10 to11, line 184 to 196
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the	Page 10, line

PRISMA 2020 Checklist

Section and Topic	Item #	Checklist item	Location where item is reported
		model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	184 to 188
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	Page 10 to 11, line 188 to 196
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	Not applicable
Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	Not applicable
Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	Not done due to highly heterogeneous data
RESULTS			
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.	Page 11, line 203 to 207 and figure 1
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	Figure 1
Study characteristics	17	Cite each included study and present its characteristics.	Page 11 to 15, line 218 to 256 and table 1
Risk of bias in studies	18	Present assessments of risk of bias for each included study.	Page 11, line 212 to 216 and supplements1,2 and 3
Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	Tables 2,3 and 4
Results of syntheses	20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies.	Page 11 to 15, line 212 to 256 and table 1, supplements1,2 and 3
	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	Page 15 to 21, line 258 to 361 and tables 2,3,4 and supplements 5,6
	20c	Present results of all investigations of possible causes of heterogeneity among study results.	Page 15 to 21, line 258 to 361 and tables 2,3,4 and supplements

PRISMA 2020 Checklist

Section and Topic	Item #	Checklist item	Location where item is reported
			5,6
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	Not applicable
Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	Not applicable
Certainty of evidence	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	Not done due to highly heterogeneous data
DISCUSSION			
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.	Page 21 to 24, line 365 to 437
	23b	Discuss any limitations of the evidence included in the review.	Pages 23 to 24, line 421 to 437
	23c	Discuss any limitations of the review processes used.	Page 24, line 439 to 448
	23d	Discuss implications of the results for practice, policy, and future research.	Page 24, line 443 to 445 and 450 to 456
OTHER INFORMATION			
Registration and protocol	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	Page 7, line 103 to 106
	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	Page 7, line 103 to 106
	24c	Describe and explain any amendments to information provided at registration or in the protocol.	None
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	None
Competing interests	26	Declare any competing interests of review authors.	None
Availability of data, code and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	None

From: Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. doi: 10.1136/bmj.n71

For more information, visit: http://www.prisma-statement.org/

BMJ Open

Minimal important change and difference for knee osteoarthritis outcome measurement tools after non-surgical interventions: a systematic review

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2022-063026.R1
Article Type:	Original research
Date Submitted by the Author:	25-Mar-2023
Complete List of Authors:	Silva, M Denika; University of Canberra, Faculty of Health; General Sir John Kotelawala Defence University, Department of physiotherapy Perriman, Diana; University of Canberra, Faculty of Health; ACT, Trauma and Orthopaedic Research Unit, Canberra Hospital Fearon, Angela; University of Canberra, Faculty of Health; Research Institute for Sport and Exercise, University of Canberra, Australia Couldrick, Jacqui; University of Canberra, Faculty of Health Scarvell, Jennie; University of Canberra, Faculty of Health; Trauma and Orthopaedic Research Unit, Canberra Hospital
Primary Subject Heading:	Rehabilitation medicine
Secondary Subject Heading:	Rehabilitation medicine
Keywords:	STATISTICS & RESEARCH METHODS, Knee < ORTHOPAEDIC & TRAUMA SURGERY, Musculoskeletal disorders < ORTHOPAEDIC & TRAUMA SURGERY

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Minimal important change and difference for knee osteoarthritis outcome measurement tools after non-surgical interventions: a systematic review

M Denika C Silva[1,2,5], Diana M Perriman[1,2,3], Angela M Fearon[1,2,4], Jacqui M Couldrick[1,2], Jennie M Scarvell[1,2]

1. Faculty of Health, University of Canberra, Australia
2. Trauma and Orthopaedic Research Unit, Canberra Hospital, Canberra, Australia
3. College of Medicine and Health Sciences, Australian National University, Canberra, Australia
4. Research Institute for Sport and Exercise, University of Canberra, Australia
5. Department of Physiotherapy, Faculty of Allied Health Sciences, General Sir John Kotelawala Defence University, Werahera, Sri Lanka

Corresponding author: M Denika C Silva (ORCID ID:0000-00018648-471X)

Address: Faculty of Health, University of Canberra, Bruce ACT 2617 Australia

Tel.: +61 (0) 433857153 **Email:** denika.silva@canberra.edu.au

Minimal important change and difference for knee osteoarthritis outcome measurement tools after non-surgical interventions: a systematic review

ABSTRACT

Objectives: To systematically review and provide estimates of the minimal important change (MIC) and difference (MID) for outcome tools in people with knee osteoarthritis (OA) after non-surgical interventions.

Design: A systematic review.

Data sources: MEDLINE, CINAHL, Web of Science, Scopus, and Cochrane databases were searched up to September 21, 2021.

Eligibility criteria: We included studies that calculated MIC and MID using any calculation method including anchor, consensus and distribution methods, for any knee OA outcome tool after non-surgical interventions.

Data extraction and synthesis: We extracted reported MIC, MID and MDC estimates. We used quality assessment tools appropriate to the studies' methods to screen out low-quality studies. Values were combined to produce a median and range, for each method.

Results: Forty-eight studies were eligible (anchor-k=12, consensus-k=1 and distribution-k=35). MIC values for 13 outcome tools including Knee injury and Osteoarthritis Outcome Score (KOOS)-pain, activities of daily living (ADL), quality of life (QOL) and Western Ontario and McMaster Universities Arthritis Index (WOMAC)-function were estimated using five high-quality anchor studies. MID values for 23 tools including KOOS-pain, ADL, QOL and WOMAC-function, stiffness and total were estimated using six high-quality anchor studies. One moderate quality consensus study reported MIC for pain, function and global assessment. Minimum detectable

change values (MDC, from distribution method estimates) for 126 tools including KOOS-QOL and WOMAC-total were estimated using 38 good-to-fair-quality studies.

Conclusion: Median MIC, MID and MDC estimates were reported for outcome tools in people with knee OA after non-surgical interventions. The results of this review clarify the current understanding of MIC, MID and MDC in the knee OA population. However, some estimates suggest considerable heterogeneity and require careful interpretation.

PROSPERO registration number: CRD42020215952

Keywords: minimal important change; minimal important difference; minimum detectable change; knee osteoarthritis

Strengths and limitations of this study

- We estimated MIC (within-group), MID (between-groups), and minimum detectable change values using anchor, consensus, or distribution methods papers respectively.
- This systematic review included a defined population of people with knee OA, after non-surgical interventions.
- High-quality anchor studies were used to contribute to MIC and MID estimates were assessed using a credibility tool specially designed to evaluate anchor method papers.
- Consensus and distribution methods papers were evaluated using quality assessment tools suited to each method.
- Median estimates were used to reflect the synthesised data due to data skewness.

INTRODUCTION

The efficacy of therapeutic interventions is commonly evaluated using statistical significance regardless of patient importance[1]. To understand whether differences in outcome measures after treatment are important to patients, it is necessary to know what constitutes a minimum important change or difference for the individual or cohort. These changes and differences are called the minimal important change (MIC) and difference (MID). There are numerous outcome measures for knee osteoarthritis (OA) and many estimates of MIC and MID. However, these estimates can arise from different methodologies leading to variability, confusion and misinterpretation[2-5]. Achieving clarity in this space is crucial as these values are used in regulatory and clinical decision making[6,7]. This systematic review aimed to provide estimates of MIC and MID for knee OA outcome measurement tools in people with knee OA after non-surgical interventions.

MIC and MID are defined as the minimum value of an outcome measure that the patient, clinician or relevant others perceive as an important change or difference[4,8,9]. The MIC considers the change in a clinical outcome measure within a single group or an individual over time. In contrast, MID considers the difference between independent groups or between individuals[4,10-12]. However, the terminology of MIC and MID is used inconsistently[13]. The concept was first described by Jaeschke, who studied patients' perceptions of pre- and post-intervention beneficial change[8]. This concept later included both improvement[14] and worsening[15].

Three methods are used to estimate MIC and MID: anchor, consensus and distribution[6,16]. For the anchor method, MIC or MID values are usually estimated by referencing the patients' responses against an externally validated scale ('anchor')[17]. The "global rating of change" is most commonly used as the anchor but other methods (proxy responses or performance based measures) are also used[18]. The receiver operating characteristic (ROC) method for deriving an estimate from anchor questions has been suggested to be more precise for clinical settings than the mean change method[9,15]. In the consensus method, values are directly estimated by a group of experienced clinicians or patients until a consensus is achieved[6]. In the distribution method, values are estimated statistically, based on the variance of the outcome data using half the standard deviation (SD)[19], one standard error of measurement (SEM)[20] or minimum detectable change (MDC) which is based on SEM[6,21]. The anchor method is widely considered to be the most valid because it is based on patient perception of what constitutes minimal change or difference[16,22]. In this paper, we have included MIC and MID estimates from anchor and consensus-based papers as well as MDC estimates. MDC estimates, as a distribution measure, are less meaningful because they do not reflect patient perception, but they do estimate instrument error which is of value to researchers[6,21]. For this reason, we included MDC as well as the anchor and consensus estimates.

Knee OA is a common cause of pain and disability[23]. Outcome measurement tools that include the domains of pain, physical function, patient global assessment and imaging are recommended to determine the efficacy of therapeutic interventions in knee OA studies[24]. MIC and MID values have been estimated for knee OA outcome measures in these domains using anchor, consensus, and distribution methods with

variable results. The variability of the methods used makes the selection of an appropriate estimate confusing for clinicians, researchers and regulatory bodies[2-4].

The primary objective of this systematic review was to estimate MIC and MID for knee OA outcome measurement tools based on estimates from high-quality anchor studies only. The secondary objectives were to determine MIC and MID estimates based on consensus method and to synthesis MDC values derived from distribution methods.

METHODS

This systematic review was designed and reported according to the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) statement[25]. The protocol was registered on PROSPERO (registration number: CRD42020215952).

Literature search

Five databases (MEDLINE, CINAHL, Web of Science, Scopus and Cochrane) were searched from each database's respective inception up to September 21, 2021. A comprehensive search strategy was developed to capture all relevant articles, and database-specific MESH terms were used. The search strategy was as follows. (*knee OR genu OR tibiofemoral OR patellofemoral) AND (osteoarthr* OR degenerat*) AND (("MCIC" OR "MCID" OR "MCII" OR "MIC" OR "MII" OR "MPCC" OR "MPCD" OR "MPCI" OR "MDC" OR "SDC" OR "SDD" OR "CIC" OR "CID") OR ("minim* clinical* important change*" OR "minim* clinical* important difference*" OR "minim* clinical*

important improvement*" OR "minim* important change*" OR "minim* important difference*" OR "minim* important improvement*" OR "minim* perceptible clinical* change*" OR "minim* perceptible clinical* difference*" OR "minim* perceptible clinical* improvement*" OR "minim* detectable change*" OR "small* detectable change* " OR "small* detectable difference* " OR "clinical* important change*" OR "clinical* important difference*")). The records were exported to EndNote version X9.2 for reference management.

Study screening and selection criteria

Covidence software (Covidence systematic review software, Veritas Health Innovation, Melbourne, Australia (www.covidence.org) was used to manage the selection process. Records identified in the search were uploaded and duplicates were removed. Screening of titles and abstracts, then full texts, were performed independently by two reviewers (DS and JC) and conflicts were resolved by a third reviewer (JS). Included studies incorporated any design that calculated MIC and MID for any knee OA outcome measurement tool considering improvement after non-surgical intervention for adults with knee OA, and using any calculation method: anchor, consensus or distribution methods. We included studies that reported MDC because MDC is considered as an estimate from the distribution method[3]. Though distribution-based approaches such as MDC do not reflect the patients perception, MDC values are important for researchers to get some idea about instrument error[6,21]. We considered studies with MIC or MID values for improvement only and excluded values for deterioration because improvement values are used to evaluate the efficacy of treatment. Studies were excluded if the data from participants with knee OA could not be separated from other conditions, e.g. hip OA or other knee pathologies. Studies

of MIC, MID and MDC were included even if they used a different terminology e.g. minimal clinically important change for MIC, minimal clinically important difference for MID and smallest detectable change or difference or minimal detectable difference for MDC.

For consistency, we defined the MIC as the pre-post change of one group i.e. threshold for those who responded that they had minimally improved on the anchor measure. The MID is defined as the difference (pre-post-change) between two groups i.e. “minimally improved” and “stayed the same” groups using the anchor response as defined in previous studies[10,12,26]. The MDC is the minimum change above the measurement error based upon a given level of confidence[6,21]. MDC values for a 90 or 95 confidence interval are labelled as MDC90 or MDC95.

Quality assessment

The quality of the included studies was assessed according to their methodology. The quality of the anchor studies was assessed using the credibility instrument developed by Devji (2020)[27] which was designed to assess the credibility of anchor studies assessing MIC and MID of patient reported outcome measures. However, we adapted this tool in the following ways. The credibility instrument includes five core criteria (1. The anchor is rated by the patient, 2.The anchor is interpretable and relevant to the patient, 3. The MIC or MID estimate of patient reported outcome measure is precise, 4.The correlation between the anchor and the outcome measure reported by the patient is satisfactory, and 5.The authors select a threshold on the anchor that reflects a small but important difference). We adapted criteria 1,3 and 4. For criteria 1 and 4, we included both patient and clinician as relevant anchor

respondents. For criteria 3, we included performance measures as well as patient reported outcome measures .We considered the paper to be “high” quality if at least three of the five criteria were “yes”, “definitely yes” or “to a great extent”; and of “low” quality if not[28]. Consensus studies were assessed using the Critical Appraisal Screening Program (CASP)-qualitative tool[29] which is designed to assess qualitative studies and is well-suited to consensus studies. The quality was rated as “high”, “moderate” or “low” based on reliability and credibility[30]. Distribution studies were evaluated using the National Heart, Lung and Blood Institute, National Institute of Health (NHLBI, NIH) quality assessment tool for before-after (pre-post) studies with no control group[31] and ratings included “good”, “fair” or “poor” based on reliability and credibility[31]. The quality assessment of included studies was performed by one reviewer (DS) and a random sample of 20% had an independent second review (AF or JC) to improve the accuracy[32].

Data extraction and analysis

We extracted study characteristics including sample size, participant demographics, details of the intervention, follow-up time, outcome measurement tools, calculation method and actual estimate reported based on the method (MIC or MID or MDC). Additionally, we extracted the details of the anchor used in each study.

We extracted reported MIC, MID and MDC values from each study. We normalised the values to a 0 to 100 scale. If a study reported MDC as a percentage of the grand mean (MDC divided by grand mean percentage)[33], we converted the data into MDC90 or 95 for the synthesis.

All data were synthesised and described as the median estimates. The median and range (minimum and maximum) of MIC, MID and MDC were calculated using multiple estimates from the included studies arising from different non-surgical interventions, calculation methods, time points, and anchors. Mean values were not calculated due to skewness of distributions[34]. We excluded low-quality anchor studies from the median MIC and MID synthesis. Furthermore, we conducted a sub-analysis to determine median MID based on the ROC method where available because the ROC estimates are considered to be more precise than mean-change estimates and recommended at both individual and group level analyses, and in clinical settings[9, 15]. Though we planned to conduct a sub-analysis to determine the effect of follow-up time on MIC and MID, we were unable to do reliable rate estimates because of a limited number of studies. Therefore, we plotted the values against time including only studies where the outcome measures were assessed at three-time points or more.

Patient and public involvement

This is a systematic review. Patients or the public were not involved in this study.

Deviations from the protocol

The protocol registered in PROSPERO lists searches in MEDLINE, Embase, CENTRAL (Cochrane Central Register of Controlled Trial), Web of Science and CINAHL. However, Embase ceased to be available to the research team, so, Scopus was substituted. The data synthesis plan to assess MIC and MID in terms of standardized mean difference (SMD) was not performed due to the skewness of the distributions. Therefore, we reported median and range for each measure without

comparison. The planned meta-analysis was prevented by the skewness and homogeneity of the data and a decision was made to follow a simple descriptive approach using median estimates that was more accessible[28].

RESULTS

Study selection

The search yielded 2376 studies and after duplicates were removed, 1059 records were screened. Two hundred and seventeen studies were screened in full-text review resulting in 48 eligible studies (k=48) **(Figure 1)**. No further studies were identified after checking the reference lists of included studies.

Included studies calculated MIC and MID by anchor method (k=12), by consensus method (k=1) and MDC by distribution method (k=35).

The methodological quality of included studies

Most anchor studies (k=10) were of high quality and two were of low quality[35,36] **(Supplement 1)**. The quality of the consensus study (k=1) was moderate **(Supplement 2)**. The quality of distribution studies ranged from good (k=11) to fair (k=24) **(Supplement 3)**.

Study characteristics of included studies

All the anchor studies were observational prospective cohort studies[14,15,35,37-43] and two of them were nested within randomised controlled trials[36,44] **(Table 1)**. The number of participants in each study ranged from 41 to 1606. The mean age and body

mass index ranged from 57.1 to 67.9 years and from 28.1 to 33 kg/m^2, respectively. The interventions used in these studies were rehabilitation, exercise, physiotherapy and non-steroidal anti-inflammatory drugs. The follow-up time ranged from seven days to one year. Eleven studies used global rating of change as the external anchor while one study[44] used multiple anchors. Most anchor studies (k=6) [36-39,42,44] reported MID values only, four studies[14,40,41,43] reported MIC values only and two studies[15,35,] reported both MIC and MID. Four of these studies[35,37,41,42] also reported MDC values. Moreover, studies used different anchor questions and group classifications when calculating MIC and MID. For example, one MIC study[35] considered the minimal improved response group as "my knee has got better" and another study[40] considered the response group as both "good" and "excellent" improvement groups **(Supplement 4)**.

Table 1: Characteristics of included studies

Study	Number of participants; Age, years, Mean (SD)	Intervention (Follow-up)	Outcome tool	Calculation method (Measure extracted)
Tubach, 2005[14]	603; 67.9 (10.2)	NSAID (4 weeks)	Pain on VAS on movement; Patient's global assessment of disease activity on a VAS; WOMAC-function subscale	Anchor (MIC)
Mills, 2016[15]	272; NR	Non-surgical (26,52 weeks)	KOOS-pain, activities of daily living, quality of life	Anchor (MIC and MID)
McCarthy, 2004[33]	15; 65.1 (11.3)	Exercises (1 week)	Aggregated locomotor function score; Eight meter walk time; Stair ascent and descent time; Transfer time	Distribution (MDC95)
Harris, 2013[35]	134; 59 (11)	Non-operative (3 months)	OKS- pain, function, summary; ICOAP; KOOS-PS	Anchor (MIC, MID); Distribution (MDC90)
Klokker, 2016[36]	41; NR	exercises (12 weeks)	DAP	Anchor (MID)
Angst, 2018[37]	190; 66.1 (10.2)	Rehabilitation (3 months)	WOMAC-pain, function, functional standing, walking and stiffness; SF-36-bodily pain, physical functioning, role-physical, vitality, social functioning, mental and general	Anchor (MID); Distribution (MDC95)
Hmamouchi, 2012[38]	173; 57.1 (10.1)	NSAID (6 weeks)	WOMAC-total	Anchor (MID)
Mostafaee, 2021[39]	142; 58.3 (7.6)	Physiotherapy (4 weeks)	KOOS-pain, symptoms, activities of daily living, sports and recreation, and quality of life	Anchor (MID)
Ornetti, 2011[40]	881; 67.1 (10.9)	NSAID (4 weeks)	WOMAC-function; Patient-reported functional disability on an NRS; Physician reported functional disability on an NRS	Anchor (MIC)
Singh, 2014[41]	137; NR	Conservative (2 weeks)	ICOAP-pain, constant pain, intermittent pain; KOOS-PS; KOOS-quality of life	Anchor (MIC); Distribution (MDC90)
Williams, 2012[42]	159; NR	Exercise (2, 6, 12 months)	KOS-activities of daily living subscale; LEFS; WOMAC-total	Anchor (MID); Distribution (MDC90, MDC95)
Perrot, 2013[43]	1606; 66.9 (9.0)	Usual care (7 days)	Pain on a 0-10 NRS at rest and movement	Anchor (MIC)
Lee, 2017[44]	165; 61	Thai Chi or Physical therapy (12 weeks)	PROMIS-Short Forms- physical function, pain interference, depression, anxiety	Anchor (MID)
Salottolo, 2018[45]	27 (clinicians); NR	NR	Pain, Function and Global assessment	Consensus (MIC)
Alghadir, 2015[46]	65; 54.9 (9.9)	NR	Timed Up and Go test	Distribution (MDC95)
Alghadir, 2016 a[47]	121; 52.9 (9.3)	NR (48 hours)	WOMAC- Arabic version-pain, function, total	Distribution (MDC95)
Alghadir, 2016 b[48]	121; 54.0 (9.3)	NR (48 hours)	NRS for pain- Arabic version	Distribution (MDC95)
Alghadir, 2017[49]	97; 58 (11.5)	NR (1 week)	OKS- Arabic version	Distribution (MDC95)
Alghadir, 2018[50]	121; 52.9 (12.5)	NR (24 hours)	VAS for pain; NRS for pain; Verbal Rating Scale for pain	Distribution (MDC95)
Suwit, 2020[51]	55; 69.0 (11)	NR (1 week)	30-seconds chair stand; 40-meter fast-paced test; 9-step climb	Distribution (MDC90)
Baert, 2018[52]	8 (12 knees); 68 (11)	NR (3 seconds)	Knee joint position sense test using analogue inclinometer; Knee force sense test using a handheld dynamometer	Distribution (MDC95)
Brisson, 2018[53]	46; 61.0 (6.6)	NR (6 months, 24 months)	Peak KAM, KAM impulse and Peak KFM using 3D motion analysis; Quadriceps strength and power using dynamometer; Load frequency using a triaxial accelerometer; BMI	Distribution (MDC95)
Callghan, 2009[54]	55; 64 (14)	NR (24- 72 hours)	Quadriceps fatigue using surface electromyography	Distribution (MDC95)

Suhail and Chaudha, 2021[55]	82; 60.4 (6.7)	NR (48 hours)	2-minute walk test	Distribution (MDC95)
Hoglund, 2019[56]	20; 58.3 (8.05)	NR (2-7 days)	30-second fast-paced walk test	Distribution (MDC95)
Ijima, 2019[57]	59; 59.1 (6.1)	NR (1 month)	Stopwatch-based 11- Step stair climb test	Distribution (MDC90, MDC95)
Jansen, 2021[58]	103; NR	NR (1 month)	Knee image: bone density, joint space width, osteophytes, eminence height, and joint angle	Distribution (MDC95)
Kean, 2010[59]	20; 53.6 (9.1)	NR (same day)	Quadriceps isokinetic strength, isometric strength and percent of voluntary activation using a dynamometer	Distribution (MDC90)
Klokker, 2015[60]	20; 64 (6.6)	NR (1 week)	DAP	Distribution (MDC95)
McCarthy, 2008[61]	55; 64.9 (9.7)	NR (1 week)	Maximum isometric strength; Vastus medialis oblique, Vastus lateralis, Rectus femoris-initial median frequency; Vastus medialis oblique, Vastus lateralis, Rectus femoris-fatigue slope	Distribution (MDC95)
Monticone, 2021[62]	102; 69.1 (9.0)	NR (10 days)	Fremantle Knee Awareness Questionnaire, Italian version	Distribution (MDC95)
Nalbant, 2021[63]	25; 62.3 (9.8)	NR (same day)	L-test	Distribution (MDC95)
Naylor, 2014[64]	75; 67.6 (9.4)	NR (10 days)	VAS for pain; Six Minute Walk Test; Timed Up and Go test; KOOS-pain, symptoms, activities of daily living, quality of life	Distribution (MDC95)
Parveen, 2017[65]	25; 50.5 (6.3)	NR (7 days)	Tinetti Performance-Oriented Mobility Assessment scale- total, balance subscale, gait subscale	Distribution (MDC95)
Peter, 2018[66]	135; NR	NR (1 week)	Animated activity questionnaire	Distribution (MDC95)
Piva, 2004[67]	25; 57 (8)	NR (same day)	Get up and Go test	Distribution (MDC95)
Pratheep, 2018[68]	20; 70.3 (5.8)	NR (2 weeks)	Pressure pain threshold using algometer	Distribution (MDC95)
Takacs, 2014[69]	20; 64.1 (7.9)	NR (14 days)	Single-leg standing balance test	Distribution (MDC95)
Tevald, 2016[70]	25; 61 (9)	NR (7 days)	Hip abductor strength using a hand-held dynamometer	Distribution (MDC95)
Tse, 2021[71]	38; 65.1 (7.0)	NR (same day)	Frontal plane tibial alignment using inclinometer	Distribution (MDC95)
Turcot, 2008[72]	25; 63.9 (7.6)	NR (8 days)	3D linear accelerations of the tibia and femur during walking	Distribution (MDC95)
Van der Straaten, 2020[73]	19; 65.1 (5.2)	NR (20 days)	Range of motion (trunk abduction-adduction, pelvis abduction-adduction, hip flexion-extension) and Center of mass displacement	Distribution (MDC95)
Yuruk, 2014[74]	40; NR	NR (One day)	De Motion Mobility Index (Turkish version)	Distribution (MDC90, MDC95)
Ravaud, 1999[75]	30; 65.4 (7.7)	NR (2 weeks)	Joint space width	Distribution (MDC95)
Kanko, 2019[76]	74; 57.7 (8.8)	Exercises (1 week)	Star excursion balance test	Distribution (MDC95)
Mutlu, 2015[77]	73 (141 knees); 56.2 (6.7)	Physiotherapy (4 weeks)	Pressure pain threshold using a handheld pressure algometer	Distribution (MDC90)
Hunter, 2006[78]	72; 58.9 (8.6)	Novel ther (6 months)	Cartilage volume using X-ray, MRI (Femur, Patella, Tibia)	Distribution (MDC95)
Motyl, 2013[79]	15; 61.0 (7.8)	Steroid injection (20 days)	20-meter walk test	Distribution (MDC95)

MIC: Minimal Important Change, MID: Minimal Important Difference, MDC90: Minimum Detectable Change based on 90% confidence interval, MDC95: Minimum Detectable Change based on 95% confidence interval, NR: Not Reported, WOMAC: Western Ontario and McMaster Universities Arthritis Index, SF-36: 36-item Short Form health survey, OKS: Oxford Knee Score, ICOAP: Intermittent and Constant Osteoarthritis Pain, KOOS-PS: Knee injury and Osteoarthritis Outcome Score Physical Function Short form, KOOS: Knee injury and Osteoarthritis Outcome Score, DAP: Dynamic weight-bearing Assessment of Pain, PROMIS: Patient-Reported Outcome Measurement Information System, NRS: numeric rating scale, VAS: Visual Analog Scale, KOS: Knee Outcome Survey, LEFS: Lower Extremity Functional Scale, KAM: knee adduction moment, 3-dimensional: 3D, KFM: knee flexion moment, MRI: Magnetic Resonance Imaging

The consensus study (k=1)[45] **(Table 1)** used a questionnaire to survey 27 clinicians from a range of specialities (orthopaedic (38%), rheumatology (33%), internal medicine (19%), and other (9%)). The clinicians were asked about MIC values for pain, function and global assessment for severe knee OA. However, participants were not asked to consider time duration nor the interventions. MIC was termed "minimal clinically important improvement" in this study.

Most distribution studies (k=30) that reported MDC, were test-retest observational studies[46-75] assessing the reliability of the outcome tool, and five used datasets from interventional cohort studies[76,77] and randomised controlled trials[33,78,79] **(Table 1)**. In the distribution studies (k=35), the number of participants in each study ranged from 8 to 135. The mean age and body mass index ranged from 50.5 to 70.3 years and from 22.7 to 35 kg/m^2, respectively. Studies estimated MDC90 and MDC95. The follow-up time ranged from the same day to one year.

The MIC estimates derived using the anchor method

The median MIC for 13 tools (with subscales) were calculated based on five high-quality anchor studies[14,15,40,41,43] using 23 estimates (**Table 2**). These estimates were based on different underlying calculations, follow-up time and anchor questions. Methods for calculating MIC included: mean change (pre and post mean change of the minimally improved group) [8,15,41], 75th centile value of the mean change of the group[40,43] and 75th centile value adjusted with the baseline score, age, and disease duration[14]. Most studies included one follow-up time (range: 7 days to 4 weeks), but one study[15] reported MIC at two time-points (26 and 52 weeks). One anchor question

was used in most studies, however, two studies[15,40] reported different MIC values based on two different anchor questions (general health status and functional state).

Table 2: Minimal Important Change (MIC) values of knee osteoarthritis outcome tools derived using the anchor method

Outcome tools	Score	Median	Minimum	Maximum	Number of estimates	Study
ICOAP-pain	100=worst	**18.5**			1	Singh, 2014[41]
ICOAP-constant pain	100=worst	**18.7**			1	Singh, 2014[42]
ICOAP-intermittent pain	100=worst	**18.4**			1	Singh, 2014[41]
KOOS-pain	100=best	**12.4**	4.3	20.1	4	Mills, 2016[15]
KOOS-activities of daily living	100=best	**8.4**	8.2	8.7	2	Mills, 2016[15]
KOOS-quality of life	100=best	**9.8**	8.0	11.6	2	Mills, 2016[15]; Singh, 2014[41]
KOOS-PS	100=worst	**2.2**			1	Singh, 2014[41]
Pain on NRS at rest	100=worst	**10.0**	10	10	2	Perrot, 2013[43]
Pain on VAS on movement	100=worst	**19.9 mm**			1	Tubach, 2005[14]
Patient-reported functional disability on an NRS	100=worst	**27.6**	27.2	27.9	2	Ornetti, 2011[40]
Patient's global assessment of disease activity on VAS	100=worst	**18.3 mm**			1	Tubach, 2005[14]
Physician-reported functional disability on an NRS	100=worst	**25.3**	25	25.5	2	Ornetti, 2011[40]
WOMAC-function	100=worst	**17.0**	9.1	17.1	3	Tubach, 2005[14]; Ornetti, 2011[40]
Estimates based on low-quality studies						
ICOAP-pain	100=worst	**13.4**			1	Harris, 2013[35]
KOOS-PS	100=worst	**12.0**			1	Harris, 2013[35]
OKS-pain	100=best	**17.3**			1	Harris, 2013[35]
OKS-function	100=best	**10.6**			1	Harris, 2013[35]
OKS-summary	100=best	**7.1**			1	Harris, 2013[35]

MIC: Minimal Important Change, ICOAP: Intermittent and constant osteoarthritis pain, KOOS: Knee injury and Osteoarthritis Outcome Score, KOOS-PS: Knee injury and Osteoarthritis Outcome Score Physical Function Short form, NRS: numeric rating scale, VAS: Visual Analog Scale, WOMAC: Western Ontario and McMaster Universities Arthritis Index, OKS: Oxford Knee Score

All the scores are from 0 to 100

The median estimates are shaded in blue
Estimates based on low-quality studies are shaded in grey

The MID estimates derived using the anchor method

The median MID for 23 tools were calculated based on six high-quality anchor studies[15,37-39,42,44,46] using 83 estimates (**Table 3**). These estimates were based on different underlying calculations, follow-up time, and anchor questions. Methods for calculating MID included: ROC method[80] only (k=3)[38,39,42], mean change (Redilmier and Lorig) method only (pre-post mean change difference between two groups)[81] (k=1)[37], both ROC and mean change methods (k=1)[15], and mean change in T-scores with multiple anchors (k=1)[44]. Most studies (k=4) included one follow-up (range: 4 to 12 weeks), but two studies included MID at multiple time-points e.g. 2, 6 and 12 months[15,42]. One anchor question was used in most studies, however, one study used multiple anchors[44].

Table 3: Minimal Important Difference (MID) values of knee osteoarthritis outcome tools derived using the anchor method

Outcome tools	Score	Median	Minimum	Maximum	Number of estimates	Study
KOOS-pain	100=best	**11.8**	4	17	13	Mills, 2016[15]; Mostafaee, 2021[39]
KOOS-activities of daily living	100=best	**2.5**	-1.5	15.5	7	Mills, 2016[15]; Mostafaee, 2021[39]
KOOS-quality of life	100=best	**6.5**	3	12.5	7	Mills, 2016[15]; Mostafaee, 2021[39]
KOOS-sports/recreation	100=best	**17.5**			1	Mostafaee, 2021[39]
KOOS-symptoms	100=best	**12.5**			1	Mostafaee, 2021[39]
KOS-activities of daily living	100=best	**6.4**	2.2	10.6	6	Williams, 2012[42]
LEFS	100=best	**6.9**	0.6	15.6	6	Williams, 2012[42]
PROMIS Short Forms-physical function*	100=worst		4.5	5.3	NR	Lee, 2017[44]
PROMIS Short Forms-pain interference*	100=worst		4.2	4.2	NR	Lee, 2017[44]
PROMIS Short Forms-depression*	100=worst		6	6.2	NR	Lee, 2017[44]
PROMIS Short Forms-anxiety*	100=worst		4.5	6.6	NR	Lee, 2017[44]
WOMAC-pain	100=best	**8.7**	7.1	21	3	Angst, 2018[37]
WOMAC-function	100=best	**14.5**	11.3	14.9	3	Angst, 2018[37]
WOMAC-stiffness	100=best	**20.2**	16.2	23.8	3	Angst, 2018[37]
WOMAC-standing/walking	100=best	**8.1**	5.9	10.2	2	Angst, 2018[37]
WOMAC-total	100=best	**6.8**	1.6	16.8	8	Williams, 2012[42]; Hmamouchi, 2012[38]
SF-36-bodily pain	100=best	**9.3**	8.2	10.4	2	Angst, 2018[37]
SF-36-physical function	100=best	**4.0**	3.8	4.2	2	Angst, 2018[37]
SF-36-role-physical	100=best	**8.5**	7.5	9.5	2	Angst, 2018[37]
SF-36-vitality	100=best	**4.5**	4.16	4.9	2	Angst, 2018[37]
SF-36-social function	100=best	**4.8**	2.6	7.0	2	Angst, 2018[37]
SF-36-mental health	100=best	**4.1**	2.9	5.2	2	Angst, 2018[37]
SF-36-general health	100=best	**6.6**	6	7.2	2	Angst, 2018[37]
Estimates based on low-quality studies						
DAP	100=worst	**24**			1	Klokker, 2016[36]
ICOAP-pain	100=worst	**7.8**			1	Harris, 2013[35]
KOOS-PS	100=worst	**7.8**			1	Harris, 2013[35]
OKS-pain	100=best	**14.3**			1	Harris, 2013[35]
OKS-function	100=best	**9.5**			1	Harris, 2013[35]
OKS-summary	100=best	**6.4**			1	Harris, 2013[35]

MID: Minimal Important Difference, OA: osteoarthritis, KOOS: Knee injury and Osteoarthritis Outcome Score, KOS: Knee Outcome Survey, LEFS: Lower Extremity Functional Scale, PROMIS: Patient-Reported Outcome Measurement Information SystemSF36: 36-item Short Form health survey, WOMAC: Western Ontario and McMaster Universities Arthritis Index, DAP: Dynamic weight-bearing Assessment of Pain, ICOAP: Intermittent and constant osteoarthritis pain,KOOS-PS: Knee injury and Osteoarthritis Outcome Score Physical Function Short form, OKS: Oxford Knee Score

All the scores are from 0 to 100
*Estimates were reported as a range
The median estimates are shaded in blue
Estimates based on low-quality studies are shaded in grey

MID estimates based on the ROC method were reported and compared with MID estimates for all methods. Four[15,38,39,42] of six studies (67%) used the ROC method. The ROC estimates were the same as the overall estimates in most cases **(Table 4).**

Table 4: Comparison of receiver operating curve (ROC) method based Minimal Important Difference (MID) estimates with overall estimates

	ROC method-based estimates, Median (range)	ROC method-based, Number of estimates (study)	Estimates regardless of calculation method, Median (range)	Number of estimates-Regardless of calculation method, (study)
KOOS-pain	**11.8 (4.0 to 18.0)**	8 (Mills, 2016[15]; Mostafaee, 2021[39])	11.8 (4.0 to 17.0)	13 (Mills, 2016[15]; Mostafaee, 2021[39])
KOOS-activities of daily living	**12 (-1.5 to 155)**	5 (Mills, 2016[15]; Mostafaee, 2021[39])	12.5 (-1.5 to 15.5)	7 (Mills, 2016[15]; Mostafaee, 2021[39])
KOOS-quality of life	**6.5 (3.0 to 12.5)**	6 (Mills, 2016[15]; Mostafaee, 2021[39])	6.5 (3.0 to 12.5)	7 (Mills, 2016[15]; Mostafaee, 2021[39])
KOOS-sports/recreation	**17.5**	1 (Mostafaee, 2021[39])	17.5	1Mostafaee, 2021[39])
KOOS-symptoms	**12.5**	1 (Mostafaee, 2021[39])	12.5	1 (Mostafaee, 2021[39])
KOS-ADL	**6.4 (2.2 to 10.6)**	6 (Williams, 2012[42])	6.4 (2.2 to 10.6)	6 (Williams, 2012[42])
LEFS	**6.9 (0.6 to 15.6)**	6 (Williams, 2012[42])	6.9 (0.6 to 15.6)	6 (Williams, 2012[42])
WOMAC-total	**7.8 (1.6 to 16.8)**	8 (Williams, 2012[42]; Hmamouchi, 2012[38])	7.8 (1.6 to 16.8)	8 (Williams, 2012[42]; Hmamouchi, 2012[38])

ROC: Receiver Operating Curve, MID: Minimal Important Difference, KOOS: Knee injury and Osteoarthritis Outcome Score, KOS: Knee Outcome Survey, LEFS: Lower Extremity Functional Scale, WOMAC: Western Ontario and McMaster Universities Arthritis Index

All the estimates are out of 100.

The median estimates are shaded in blue
Shaded in green: Estimates of these MID values were based on the ROC method only

The effect of follow-up time on MIC and MID

There were insufficient data to establish reliable rate estimates for the effect of time. The MIC of Knee injury and Osteoarthritis Outcome Score (KOOS)-pain and

KOOS-quality of life (QOL) and, the MID of KOOS-pain, KOOS-QOL, Knee Outcome Score (KOS)-activity of daily living (ADL), Lower Extremity Functional Scale (LEFS) and Western Ontario and McMaster Universities Arthritis Index (WOMAC)-total were assessed at more than three different time points. The MIC of KOOS-QOL and, MID of KOOS-QOL, LEFS and KOS-ADL appeared to increase with increasing follow-up time. However, MIC of KOOS-pain and MID of WOMAC-total appeared to reduce with follow-up time and KOOS-pain remained constant (**Supplement 5).**

MIC values derived using the consensus method

One consensus study[45] reported that MIC for pain, function and global assessment were 20% of the maximum score.

The MDC estimates derived using the distribution method

The median MDC was calculated for 126 tools based on 38 studies (35 good-to-fair distribution and three high-quality anchor studies) using 308 estimates (**Supplement 6)**. These estimates were based on different calculation methods and follow-up times. Four included studies reported MDC90 values only[41,51,59,77] and 29 studies reported MDC95 values only. Five studies[37,42,57,64,74] reported both the MDC90 and MDC95 values. Most studies (k=37) reported unadjusted MDC, while one study reported both the adjusted and unadjusted estimates[37]. Six studies separately reported inter/intra-rater MDC values[46,52,56,60,67,73]. Furthermore, three studies reported distinct values for two patient groups in each study, for example, the placebo group and the treatment group[78], the most painful and the least painful groups[70], and the groups that reported moderate improvement ("great deal better") and MCID improvement ("somewhat better")[41]. Most studies assessed the index (worst) knee,

but one study based the estimate on all diseased knees (12 knees in 8 patients)[52]. Regarding the time point of MDC estimation, most studies (k=39) reported MDC estimates at one time-point only, but one study reported MDC estimates at three-time points (2, 6 and 12 months)[42].

DISCUSSION

This systematic review provided estimates for MIC and MID of knee OA outcome tools after non-surgical interventions derived using anchor, consensus and distribution methods respectively. This review is unique in that it provides estimates for MIC, MID (based on high-quality studies) and MDC (from good to fair-quality studies) of knee OA outcome tools after non-surgical interventions. MDC was reported for a greater number of outcome measures (126) than for MIC (13) or MID (23). MID estimates based on the ROC method were similar to the overall median estimates, however, the majority of MID studies used the ROC method. Although we found that some MIC and MID appear to increase with follow-up time, this was not consistent.

The estimates for MIC and MID reported in this review are lower than those reported previously[82-84]. Previous reviews which included knee replacement interventions[82-84] produced higher estimates suggesting that knee replacement cohorts need more improvement to be satisfied. The MID values for WOMAC-pain and function in this review ranged from 7.1 to 21 and 11.3 to 14.9 (out of 100) respectively; compared to reviews of total knee replacements which reported values ranging from 4.0 to 47.9 and 1.8 to 33.0 (out of 100)[82-84]. This disparity may be due to differences in disease severity which has been previously reported based on baseline pain

score[6,15,85]. Therefore, our data is more applicable to patients and cohorts receiving non-surgical interventions. Furthermore, previous knee OA intervention studies have used MID estimates from studies with combined hip and knee OA[86-87]. Given that MID is sensitive to disease type[6], our median estimates are likely to be more applicable to the knee OA population.

Some of the median estimates presented in this study suggest considerable heterogeneity. For example, the MID for WOMAC-pain was 8.7, but, the range extended from 7.1 to 21. These wide ranges are seen for other estimates including MIC for KOOS-pain, KOOS-QOL and LEFS. The median estimate was used because it is robust when data is skewed. However, the uncertainty which accompanies the wide ranges reported must be acknowledged.

MDC was reported for more outcome measures than MIC or MID. MDC is derived from data distribution only, unlike MIC and MID which are related to patients' perception[17,22]. Researchers may use MDC estimates as an option if MIC or MID are not reported. Yet, according to the results of this study and others, MDC can be larger or smaller than MID[4,10]. Hence, researchers using MDC estimates from single studies to establish a sample size may over-or under-estimate the number of participants required for a given power.

The ROC estimates were similar to the synthesised MID estimates which used all calculation methods. Our synthesised estimates were based on a combination of both the mean change[81] and ROC methods[80]. However, the ROC estimates are reported to be more precise, and can be applied to both individuals and groups and

are recommended in clinical settings[9,15]. Moreover, the area under the curve of the ROC has the advantage of being able to interpret the level of confidence for the MID estimate from acceptable to outstanding discrimination between responders and non-responders[17]. Therefore, we recommend using our median ROC based MID estimates where possible.

Although we found that some MIC and MID appear to increase with follow-up time, this was not consistent. Two previous studies[15,42] suggested that there may be an effect of time, due to changing of perceptions over time (response-shift), especially in patients with chronic conditions[15,88]. Additionally, recall bias is affected by increased follow-up time and may also affect estimates[6,27,89]. Therefore, although the consistency of follow-up time must be considered, more data is required to determine the effect of follow up time on MIC and MID.

One of the studies included in this review used the consensus method. They reported that MIC was 20% of the maximum score for pain, function and global assessment[45], but, our anchor studies data suggest that MIC is highly variable (2.2 to 27.6 out of 100) depending on the outcome measurement. Therefore, the blanket application of 20% may not be suggested regardless of the tool used.

In this review, we considered only MIC, MID and MDC but there are other measures of clinical improvement. While the MIC and MID are used to assess meaningful clinical effects, recent reports have questioned the applicability of these concepts as they do not consider the costs, risks, benefits and inconvenience of the treatment. The smallest worthwhile effect (SWE) was developed using the benefit-

harm trade-off method, described by Barrett in 2005[90]. The SWE is defined as the smallest amount of improvement which is identified by the patient as worthwhile when considering the improvement outweighing risks and inconvenience[91] and the estimates are always compared to natural recovery[92]. However, only one study has reported SWE for people undergoing total knee replacement[93]. Other studies have evaluated "patients acceptable symptom state" (PASS) which is the symptom state that patients consider acceptable or when they feel "well" after treatment[84,94,95]. PASS estimates for WOMAC function are reported to be between 31 and 34.4[96]. These values are much higher than our MIC median estimate of 17 (9.1 to 17.1). Although MIC and MID are still commonly used, the development of this field of research will enable value judgements as well as clinical judgements to be considered in the interpretation of clinical trials of interventions.

This systematic review should be considered in light of its limitations. The results of this review have been affected by heterogeneity of the included studies including: sample size, participant demographics, severity of knee OA, varied interventions, follow-up time and calculation methods. Median estimates were used because of the data skewness, but some of the ranges were wide, challenging the certainty of some of these estimates. The reader is encouraged to take the range of the data into account when interpreting the results. Previous evidence suggests that data from follow-up times of less than one month are more reliable[27,97,98]. However, we included all estimates regardless of follow-up time. Statistical analysis was not conducted to determine the effect of follow-up time due to limited available data, but this is an interesting area for further study. The grey literature was not searched for this review.

This review presents median estimates for MIC, MID and MDC of people with knee OA following non-surgical interventions. A subset of MID estimates based on the ROC method is reported and, where available, this estimate is recommended as the most precise for both individual and group analyses and clinical settings. MDC estimates are available for more outcome measures but are purely statistical and arguably less applicable. This review clarifies the current understanding of MIC, MID and MDC in the knee OA population. However, some estimates suggest considerable heterogeneity and require careful interpretation.

Author contributions

MDCS, DMP, AMF, and JMS conceptualised and designed the study. MDCS ran the database search. MDCS and JMC did the screening, full-text review and data acquisition. MDCS, DMP, AMF and JMS were involved in data analysis and Interpretation of the data. MDCS wrote the first draft of the manuscript. All authors reviewed the draft, read and approved the final manuscript.

Role of the funding source

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Conflict of interest

All the authors declare that there are no conflicts of interest in this work.

Ethics approval

This study is a systematic review and does not involve human participants. Therefore, ethics approval is not applicable.

Data Sharing

All data relevant to the study are included in the article or uploaded as supplementary information.

REFERENCES

1. Harris JD, Brand JC, Cote MP, Faucett SC, Dhawan A. Research Pearls: The significance of statistics and perils of pooling. part 1: clinical versus statistical significance. Arthroscopy 2017; 33: 1102-1112.
2. Terwee CB, Roorda LD, Dekker J, Bierma-Zeinstra SM, Peat G, Jordan KP, et al. Mind the MIC: large variation among populations and methods. J Clin Epidemiol 2010; 63: 524-534.
3. de Vet HC, Terwee CB, Ostelo RW, Beckerman H, Knol DL, Bouter LM. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. Health Qual Life Outcomes 2006; 4: 54-54.
4. de Vet HC, Terwee CB. The minimal detectable change should not replace the minimal important difference. J Clin Epidemiol 2010; 63: 804-805; author reply 806.
5. de Vet HCW, Beckerman H, Terwee CB, Terluin B, Bouter LM. Definition of clinical differences. J Rheumatol 2006; 33: 434.
6. Katz NP, Paillard FC, Ekman E. Determining the clinical importance of treatment benefits for interventions for painful orthopedic conditions. J Orthop Surg Res 2015; 10: 24.

7. Wright A, Hannon J, Hegedus EJ, Kavchak AE. Clinimetrics corner: a closer look at the minimal clinically important difference (MCID). J Man Manip Ther 2012; 20: 160-166.

8. Jaeschke R, Singer J, Guyatt GH. Measurement of health status: Ascertaining the minimal clinically important difference. Control Clin Trials 1989; 10: 407-415.

9. Wells G, Anderson J, Beaton D, Bellamy N, Boers M, Bombardier C, et al. Minimal clinically important difference module: summary, recommendations, and research agenda. J Rheumatol 2001; 28: 452-454.

10. Beard DJ, Harris K, Dawson J, Doll H, Murray DW, Carr AJ, et al. Meaningful changes for the Oxford hip and knee scores after joint replacement surgery. J Clin Epidemiol 2015; 68: 73-79.

11. King MT, Dueck AC, Revicki DA. Can methods developed for interpreting group-level patient-reported outcome data be applied to individual patient management? Med Care 2019; 57.

12. Clement ND, Bardgett M, Weir D, Holland J, Gerrand C, Deehan DJ. What is the minimum clinically important difference for the WOMAC index after TKA? Clin Orthop Relat Res 2018; 476: 2005-2014.

13. King MT. A point of minimal important difference (MID): a critique of terminology and methods. Expert Rev Pharmacoecon Outcomes Res 2014; 11: 171-184.

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

15. Mills KA, Naylor JM, Eyles JP, Roos EM, Hunter DJ. Examining the minimal important difference of patient-reported outcome measures for individuals with knee osteoarthritis: A model using the Knee Injury and Osteoarthritis Outcome Score. J Rheumatol 2016; 43: 395-404.

16. Make B. How can we assess outcomes of clinical trials: The MCID approach. COPD: J Chronic Obstr Pulm Dis 2007; 4: 191-194.

17. Copay AG, Subach BR, Glassman SD, Polly DW, Jr., Schuler TC. Understanding the minimum clinically important difference: a review of concepts and methods. Spine J 2007; 7: 541-546.

18. Carrasco-Labra A, Devji T, Qasim A, Phillips MR, Wang Y, Johnston BC, Devasenapathy N, Zeraatkar D, Bhatt M, Jin X, Brignardello-Petersen R. Minimal important difference estimates for patient-reported outcomes: a systematic survey. Journal of Clinical Epidemiology. 2021; 133:61-71.

19. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Med Care 2003; 41: 582-592.

20. Wyrwich KW. Minimal important difference thresholds and the standard error of measurement: is there a connection? J. Biopharm. Stat. 2004; 14: 97-110.

21. Martin Bland J, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 327: 307-310.

22. Angst F, Aeschlimann A, Angst J. The minimal clinically important difference raised the significance of outcome effects above the statistical level, with methodological implications for future studies. J Clin Epidemiol 2017; 82: 128-136.

23. Cross M, Smith E, Hoy D, Nolte S, Ackerman I, Fransen M, et al. The global burden of hip and knee osteoarthritis: estimates from the global burden of disease 2010 study. Ann Rheum Dis 2014; 73: 1323.

24. Bellamy N, Kirwan J, Boers M, Brooks P, Strand V, Tugwell P, et al. Recommendations for a core set of outcome measures for future phase III clinical trials in knee, hip, and hand osteoarthritis. Consensus development at OMERACT III. J Rheumatol 1997; 24: 799-802.

25. Moher D, Liberati A, Tetzlaff J, Altman DG, Group atP. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. Ann Intern Med 2009; 151: 264-269.

26. Dawson J, Boller I, Doll H, et al. Minimally important change was estimated for the Manchester-Oxford Foot Questionnaire after foot/ankle surgery. J Clin Epidemiol 2014;67(6):697-705.

27. Devji T, Carrasco-Labra A, Qasim A, Phillips M, Johnston BC, Devasenapathy N, et al. Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study. BMJ 2020; 369: m1714.

28. Hao Q, Devji T, Zeraatkar D, Wang Y, Qasim A, Siemieniuk RAC, et al. Minimal important differences for improvement in shoulder condition patient-reported outcomes: a systematic review to inform a BMJ Rapid Recommendation. BMJ Open 2019; 9: e028777.

29. Programme CAS. CASP (Qualitative) Checklist. [online] 2018.

30. Long HA, French DP, Brooks JM. Optimising the value of the critical appraisal skills programme (CASP) tool for quality appraisal in qualitative evidence synthesis. Research Methods in Medicine & Health Sciences 2020; 1: 31-42.

31. NHLBI N-NH, Lung, and Blood Institute, National Institute of Health quality assessment tool for before-after (pre-post) studies with no control group. 2013.
32. Haby MM, Chapman E, Clark R, Barreto J, Reveiz L, Lavis JN. What are the best methodologies for rapid reviews of the research evidence for evidence-informed decision making in health policy and practice: a rapid review. Health Res Policy Syst 2016; 14: 83.
33. McCarthy CJ, Oldham JA. The reliability, validity and responsiveness of an aggregated locomotor function (ALF) score in patients with osteoarthritis of the knee. Rheumatology (Oxford) 2004; 43: 514-517.
34. Higgins JPT CJ, Cumpston M, Li T, Page MJ, Welch VA. Cochrane handbook for systematic reviews of interventions. version 6.0 edition2019.
35. 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; 3: e003365.
36. Klokker L, Christensen R, Waehrens EE, Bandak E, Bartholdy C, Bliddal H, et al. Dynamic weight-bearing assessment of pain in knee osteoarthritis: construct validity, responsiveness, and interpretability in a research setting. Health Qual. Life Outcomes 2016; 14: 91.
37. Angst F, Benz T, Lehmann S, Aeschlimann A, Angst J. Multidimensional minimal clinically important differences in knee osteoarthritis after comprehensive rehabilitation: a prospective evaluation from the Bad Zurzach Osteoarthritis Study. RMD Open 2018; 4: e000685.
38. Hmamouchi I, Allali F, Tahiri L, et al. Clinically important improvement in the WOMAC and predictor factors for response to non-specific non-steroidal anti-

627 inflammatory drugs in osteoarthritic patients: a prospective study. BMC Res
628 Notes 2012;5:58
629 39. Mostafaee N, Nourollahi F, Mostamand J, Negahban H. Responsiveness and
630 the minimal important change of Knee injury and Osteoarthritis Outcome Score
631 in Persian patients with knee osteoarthritis following physiotherapy intervention.
632 Physiother Theory Pract 2021: 1-10.
633 40. Ornetti P, Dougados M, Paternotte S, Logeart I, Gossec L. Validation of a
634 numerical rating scale to assess functional impairment in hip and knee
635 osteoarthritis: comparison with the WOMAC function scale. Ann Rheum Dis
636 2011; 70: 740-746.
637 41. Singh JA, Luo R, Landon GC, Suarez-Almazor M. Reliability and clinically
638 important improvement thresholds for osteoarthritis pain and function scales: a
639 multicenter study. J Rheumatol 2014; 41: 509-515.
640 42. Williams VJ, Piva SR, Irrgang JJ, Crossley C, Fitzgerald GK. Comparison of
641 reliability and responsiveness of patient-reported clinical outcome measures in
642 knee osteoarthritis rehabilitation. J Orthop Sports Phys Ther 2012; 42: 716-723.
643 43. Perrot S, Bertin P. "Feeling better" or "feeling well" in usual care of hip and knee
644 osteoarthritis pain: determination of cutoff points for patient acceptable
645 symptom state (PASS) and minimal clinically important improvement (MCII) at
646 rest and on movement in a national multicenter cohort study of 2414 patients
647 with painful osteoarthritis. J Pain 2013; 154: 248-256.
648 44. Lee AC, Driban JB, Price LL, Harvey WF, Rodday AM, Wang C.
649 responsiveness and minimally important differences for 4 Patient-Reported
650 Outcomes Measurement Information System short forms: physical function,

pain interference, depression, and anxiety in knee osteoarthritis. J Pain 2017; 18: 1096-1110.

45. Salottolo K, Stahl E. Minimal clinically important improvement response in patients with severe osteoarthritis of the knee: Short report from a survey of clinicians. J Orthop 2018; 15: 424-425.

46. Alghadir A, Anwer S, Brismée J-M. The reliability and minimal detectable change of Timed Up and Go test in individuals with grade 1-3 knee osteoarthritis. BMC Musculoskelet Disord 2015; 16: 174-174.

47. Alghadir A, Anwer S, Iqbal ZA, Alsanawi HA. Cross-cultural adaptation, reliability and validity of the Arabic version of the reduced Western Ontario and McMaster Universities Osteoarthritis index in patients with knee osteoarthritis. Disabil Rehabil 2016; 38: 689-694.

48. Alghadir AH, Anwer S, Iqbal ZA. The psychometric properties of an Arabic numeric pain rating scale for measuring osteoarthritis knee pain. Disabil Rehabil 2016; 38: 2392-2397.

49. Alghadir AH, Al-Eisa ES, Anwer S. Cross-cultural adaptation and psychometric analysis of the Arabic version of the oxford knee score in adult male with knee osteoarthritis. BMC Musculoskelet Disord 2017; 18: 190.

50. Alghadir AH, Anwer S, Iqbal A, Iqbal ZA. Test-retest reliability, validity, and minimum detectable change of visual analog, numerical rating, and verbal rating scales for measurement of osteoarthritic knee pain. J Pain Res 2018; 11: 851-856.

51. Suwit A, Rungtiwa K, Nipaporn T. Reliability and validity of the Osteoarthritis Research Society International minimal core set of recommended

performance-based tests of physical function in knee osteoarthritis in community-dwelling adults. Malays J Med Sci 2020; 27: 77-89.

52. Baert IAC, Lluch E, Struyf T, Peeters G, Van Oosterwijck S, Tuynman J, et al. Inter- and intrarater reliability of two proprioception tests using clinical applicable measurement tools in subjects with and without knee osteoarthritis. Musculoskelet Sci Pract 2018; 35: 105-109.

53. Brisson NM, Stratford PW, Maly MR. Relative and absolute test-retest reliabilities of biomechanical risk factors for knee osteoarthritis progression: benchmarks for meaningful change. Osteoarthr Cartil 2018; 26: 220-226.

54. Callaghan MJ, McCarthy CJ, Oldham JA. The reliability of surface electromyography to assess quadriceps fatigue during multi joint tasks in healthy and painful knees. J Electromyogr Kinesiol 2009; 19: 172-180.

55. Suhail A, Chaudhary S. Test-retest reliability and minimum detectable change of 2-minute walk test among individuals with knee osteoarthritis. J Clin Diagnostic Res 2021; 15: 4-7.

56. Hoglund LT, Folkins E, Pontiggia L, Knapp MW. The validity, reliability, measurement error, and minimum detectable change of the 30-second fast-paced walk test in persons with knee osteoarthritis: a novel test of short-distance walking ability. ACR Open Rheumatol 2019; 1: 279-286.

57. Iijima H, Shimoura K, Eguchi R, Aoyama T, Takahashi M. Concurrent validity and measurement error of stair climb test in people with pre-radiographic to mild knee osteoarthritis. Gait Posture 2019; 68: 335-339.

58. Jansen MP, Welsing PMJ, Vincken KL, Mastbergen SC. Performance of knee image digital analysis of radiographs of patients with end-stage knee osteoarthritis. Osteoarthr Cartil 2021; 29: 1530-1539.

59. Kean CO, Birmingham TB, Garland SJ, Bryant DM, Giffin JR. Minimal detectable change in quadriceps strength and voluntary muscle activation in patients with knee osteoarthritis. Arch Phys Med Rehabil 2010; 91: 1447-1451.

60. Klokker L, Christensen R, Osborne R, Ginnerup E, Waehrens EE, Bliddal H, et al. Dynamic weight-bearing assessment of pain in knee osteoarthritis: a reliability and agreement study. Qual Life Res 2015; 24: 2985-2992.

61. McCarthy CJ, Callaghan MJ, Oldham JA. The reliability of isometric strength and fatigue measures in patients with knee osteoarthritis. Man Ther 2008; 13: 159-164.

62. Monticone M, Sconza C, Portoghese I, Nishigami T, Wand BM, Sorrentino G, et al. Cross-cultural adaptation, reliability and validity of the Fremantle Knee Awareness Questionnaire in Italian subjects with painful knee osteoarthritis. Health Qual Life Outcomes 2021; 19: 114.

63. Nalbant A, Unver B, Karatosun V. Test–retest reliability of the L-test in patients with advanced knee osteoarthritis. Physiother Theory Pract 2021: 1-5.

64. Naylor J M, Hayen A, Davidson E, Hackett D, Harris I A, Kamalasena G, et al. Minimal detectable change for mobility and patient-reported tools in people with osteoarthritis awaiting arthroplasty. BMC Musculoskelet Disord 2014: 15, 235.

65. Parveen H, Noohu MM. Evaluation of psychometric properties of Tinetti performance-oriented mobility assessment scale in subjects with knee osteoarthritis. Hong Kong Physiother J 2017; 36: 25-32.

66. Peter WF, de Vet HCW, Terwee CB. Reliability of the Animated Activity Questionnaire for assessing activity limitations of patients with hip and knee osteoarthritis. Musculoskeletal Care 2018; 16: 363-369.

67. Piva SR, Fitzgerald GK, Irrgang JJ, Bouzubar F, Starz TW. Get up and go test in patients with knee osteoarthritis. Arch Phys Med Rehabil. 2004;85(2):284-289.

68. Srimurugan Pratheep N, Madeleine P, Arendt-Nielsen L. Relative and absolute test-retest reliabilities of pressure pain threshold in patients with knee osteoarthritis. Scand J Pain 2018; 18: 229-236.

69. Takacs J, Carpenter MG, Garland SJ, Hunt MA. Test re-test reliability of centre of pressure measures during standing balance in individuals with knee osteoarthritis. Gait Posture 2014; 40: 270-273.

70. Tevald MA, Murray A, Luc BA, Lai K, Sohn D, Pietrosimone B. Hip abductor strength in people with knee osteoarthritis: A cross-sectional study of reliability and association with function. Knee 2016; 23: 57-62.

71. Tse CTF, Charlton JM, Lam J, Ho J, Bears J, Serek A, et al. Smartphone inclinometry is a valid and reliable tool for measuring frontal plane tibial alignment in healthy and osteoarthritic knees. Phys Ther 2021; 101: pzab091.

72. Turcot K, Aissaoui R, Boivin K, Hagemeister N, Pelletier M, de Guise JA. Test-retest reliability and minimal clinical change determination for 3-dimensional tibial and femoral accelerations during treadmill walking in knee osteoarthritis patients. Arch Phys Med Rehabil 2008; 89: 732-737.

73. van der Straaten R, Wesseling M, Jonkers I, Vanwanseele B, Bruijnes A, Malcorps J, et al. Discriminant validity of 3D joint kinematics and centre of mass displacement measured by inertial sensor technology during the unipodal stance task. PLoS One 2020; 15: e0232513.

74. Yuruk Z O, Aytar A, Tüzün E H, Eker L, Yüksel I, Morton N A. Acceptability, reliability and validity of the turkish version of the De morton mobility index in elderly patients with knee osteoarthritis. Turk Geriatri Dergisi 2014; 17.

75. Ravaud P, Giraudeau B, Auleley GR, Edouard-Noël R, Dougados M, Chastang C. Assessing smallest detectable change over time in continuous structural outcome measures: application to radiological change in knee osteoarthritis. J Clin Epidemiol 1999; 52: 1225-1230.

76. Kanko LE, Birmingham TB, Bryant DM, Gillanders K, Lemmon K, Chan R, et al. The star excursion balance test is a reliable and valid outcome measure for patients with knee osteoarthritis. Osteoarthritis Cartilage 2019; 27: 580-585.

77. Mutlu EK, Ozdincler AR. Reliability and responsiveness of algometry for measuring pressure pain threshold in patients with knee osteoarthritis. J Phys Ther Sci. 2015;27(6):1961-1965.

78. Hunter DJ, Conaghan PG, Peterfy CG, Bloch D, Guermazi A, Woodworth T, et al. Responsiveness, effect size, and smallest detectable difference of Magnetic Resonance Imaging in knee osteoarthritis. Osteoarthritis Cartilage 2006; 14 Suppl A: A112-115.

79. Motyl JM, Driban JB, McAdams E, Price LL, McAlindon TE. Test-retest reliability and sensitivity of the 20-meter walk test among patients with knee osteoarthritis. BMC Musculoskelet Disord 2013; 14: 166

80. Perkins NJ, Schisterman EF. The Inconsistency of “optimal” cutpoints obtained using two criteria based on the receiver operating characteristic curve. Am J Epidemiol 2006; 163: 670-675.

81. Redelmeier DA, Lorig K. Assessing the clinical importance of symptomatic improvements. An illustration in rheumatology. Arch Intern Med 1993; 153: 1337-1342.

82. Celik D, Coban O, Kilicoglu O. Minimal clinically important difference of commonly used hip-, knee-, foot-, and ankle-specific questionnaires: a systematic review. J Clin Epidemiol 2019; 113: 44-57.

83. Maredupaka S, Meshram P, Chatte M, Kim WH, Kim TK. Minimal clinically important difference of commonly used patient-reported outcome measures in total knee arthroplasty: review of terminologies, methods and proposed values. Knee Surg Relat Res 2020; 32: 19.

84. MacKay C, Clements N, Wong R, Davis AM. A systematic review of estimates of the minimal clinically important difference and patient acceptable symptom state of the Western Ontario and McMaster Universities Osteoarthritis Index in patients who underwent total hip and total knee replacement. Osteoarthritis Cartilage 2019; 27: 1408-1419.

85. Escobar A, Riddle DL. Concordance between important change and acceptable symptom state following knee arthroplasty: the role of baseline scores. Osteoarthritis Cartilage 2014; 22: 1107-1110.

86. Angst F, Aeschlimann A, Steiner W, Stucki G. Responsiveness of the WOMAC osteoarthritis index as compared with the SF-36 in patients with osteoarthritis of the legs undergoing a comprehensive rehabilitation intervention. Ann Rheum Dis 2001; 60: 834-840.

87. Bellamy N. Pain assessment in osteoarthritis: experience with the WOMAC osteoarthritis index. Semin. Arthritis Rheum 1989; 18: 14-17.

88. Schwartz CE, Andresen EM, Nosek MA, Krahn GL. Response shift theory: important implications for measuring quality of life in people with disability. Arch Phys Med Rehabil 2007; 88: 529-536.

89. Guyatt GH, Norman GR, Juniper EF, Griffith LE. A critical look at transition ratings. J Clin Epidemiol 2002; 55: 900-908.

90. Barrett B, Brown D, Mundt M, Brown R. Sufficiently important difference: expanding the framework of clinical significance. Med Decis Making 2005; 25: 250-261.

91. Ferreira M. Research Note: The smallest worthwhile effect of a health intervention. J Physiother 2018; 64: 272-274.

92. Ferreira ML, Herbert RD, Ferreira PH, Latimer J, Ostelo RW, Nascimento DP, et al. A critical review of methods used to determine the smallest worthwhile effect of interventions for low back pain. J Clin Epidemiol 2012; 65: 253-261.

93. Wofford N. Smallest worthwhile effect values for pain and function after a total knee replacement. health related sciences, vol. PhD: Virginia Commonwealth University 2020.

94. Bellamy N, Hochberg M, Tubach F, Martin-Mola E, Awada H, Bombardier C, et al. Development of multinational definitions of minimal clinically important improvement and patient acceptable symptomatic state in osteoarthritis. Arthritis Care Res (Hoboken) 2015; 67: 972-980.

95. Kvien TK, Heiberg T, Hagen KB. Minimal clinically important improvement/difference (MCII/MCID) and patient acceptable symptom state (PASS): what do these concepts mean?. Ann Rheum Dis. 2007;66 Suppl 3(Suppl 3):iii40-iii41. doi:10.1136/ard.2007.079798

96. Tubach F, Ravaud P, Baron G, Falissard B, Logeart I, Bellamy N, et al. Evaluation of clinically relevant states in patient reported outcomes in knee and hip osteoarthritis: the patient acceptable symptom state. Ann Rheum Dis 2005; 64: 34-37.

97. Turner D, Schünemann HJ, Griffith LE, Beaton DE, Griffiths AM, Critch JN, et al. Using the entire cohort in the receiver operating characteristic analysis maximizes precision of the minimal important difference. J Clin Epidemiol 2009; 62: 374-379.

98. Lurie F, Kistner RL. In prospective study using Specific Quality of Life & Outcomes Response-Venous (SQOR-V) questionnaire the recall bias had the same magnitude as the minimally important difference. Qual Life Res 2011; 20: 1589-1593.

List of tables

Figure legends

Figure 1: Flow diagram of study selection (PRISMA, Preferred Reporting Items for Systematic reviews and Meta-Analyses)[25]

List of Supplements

Supplement 1: Quality of anchor studies assessed using the Devji (2020) credibility instrument[27]

Supplement 2: Quality of consensus studies assessed using the Critical Appraisal Screening Program (CASP)-qualitative checklist[29]

Supplement 3: Quality of distribution studies assessed using the National Institute of Health (NIH) quality assessment tool for before-after (pre-post) studies with no control group[31]

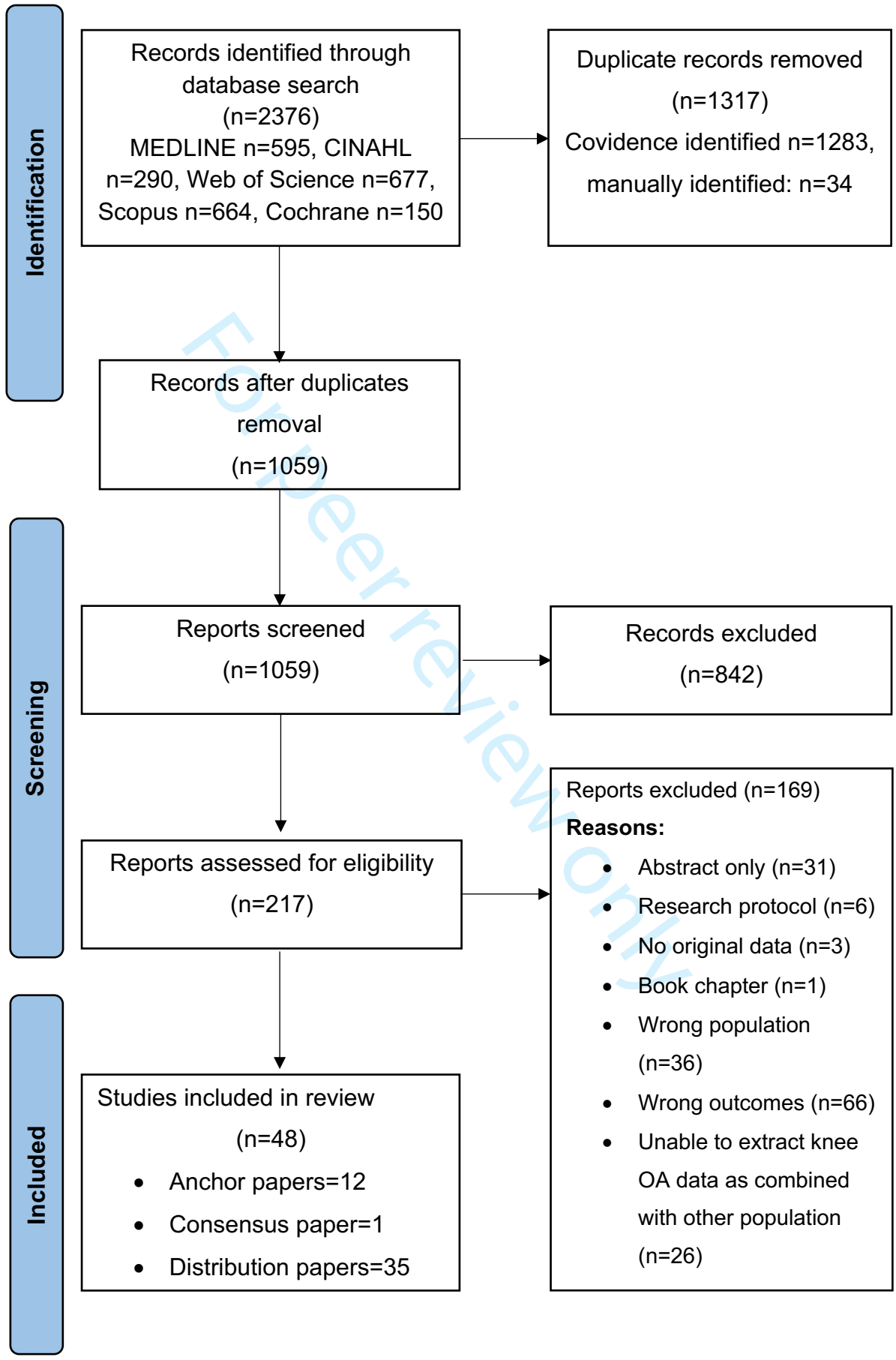
Supplement 4: Anchor properties of included anchor studies

Supplement 5: The effect of follow-up time on **A.** Minimal Important Change (MIC) and **B.** Minimal Important Difference (MID)

†: multiple MIC estimates using two different anchor questions; *: multiple MID estimates using two different anchor questions and different calculation methods

KOOS: Knee injury and Osteoarthritis Outcome Score, QOL: Quality of Life, KOS: Knee Outcome Survey, ADL: Activities of Daily Living, LEFS: Lower Extremity Functional Scale, WOMAC: Western Ontario and McMaster Universities Arthritis Index

Supplement 6: Minimum Detectable Change (MDC) values of knee osteoarthritis outcome tools derived using the distribution method



Supplement 1: Quality of anchor studies assessed using the Devji (2020) credibility instrument

Study	Item 1*	Item 2#	*Item 3#*	Item 4#	Item 5#	The overall quality of the studies**
Angst, 2018	Yes	Impossible to tell	Definitely No	To a greater extent	To a great extent	**High**
Harris, 2013	Yes	To a great extent	Impossible to tell	Not so much	Definitely No	**Low**
Hmamouchi, 2012	Yes	Not so much	Impossible to tell	To a greater extent	To a greater extent	**High**
Klokker, 2016	Yes	To a great extent	Impossible to tell	Definitely No	Not so much	**Low**
Lee, 2017	Yes	Definitely Yes	Impossible to tell	To a greater extent	Impossible to tell	**High**
Mills, 2016	Yes	Definitely Yes	Not so much	To a greater extent	Definitely Yes	**High**
Mostafaee, 2021	Yes	To a great extent	To a great extent	To a great extent	Not so much	**High**
Ornetti, 2011	Yes	To a great extent	Impossible to tell	To a greater extent	Definitely No	**High**
Perrot, 2013	Yes	To a great extent	Not so much	Definitely Yes	Not so much	**High**
Singh, 2014	Yes	To a great extent	Impossible to tell	Not so much	To a great extent	**High**
Tubach, 2005	Yes	To a great extent	Impossible to tell	Definitely Yes	Not so much	**High**
Williams, 2012	Yes	Not so much	Impossible to tell	To a greater extent	Definitely Yes	**High**

Item 1: Is the patient or necessary proxy responding directly to both the PROM and the anchor?
Item 2: Is the anchor easily understandable and relevant for patients or a necessary proxy?
Item 3: Has the anchor shown a good correlation with the PROM?
Item 4: Is the MID precise?
Item 5: Does the threshold or difference between groups on the anchor used to estimate the MID reflect a small but important difference?

* The responses to items: Yes, No, Impossible to tell
# The responses to items: Definitely yes, To a great extent, Not so much, Definitely no, Impossible to tell
** overall quality: three of the five criteria were met “Yes” or “definitely yes” or “to a great extent”, the paper is of “high” quality If not, that the paper is of “low” quality[27]

Low quality (credibility) studies are shaded.

PROM: Patient-Reported Outcome Measure, MID-Minimal Important Difference (In this study minimal important change/difference)

Supplement 2: Quality of consensus studies assessed using the Critical Appraisal Screening Program (CASP)-qualitative checklist

Study (Author, Year)	Item 1 #	Item 2 #	Item 3 #	Item 4 #	Item 5 #	Item 6 #	Item 7 #	Item 8	Item 9 #	Item 10	The overall quality of the studies µ
Salottolo, 2018	Yes	Yes	Can't tell	Can't tell	Yes	Yes	Can't tell	Yes	Yes	Researcher discusses the the contribution the study makes to existing knowledge or understanding	**Moderate**

Item 1: Was there a clear statement of the aims of the research?
Item 2: Is a qualitative methodology appropriate?
Item 3: Was the research design appropriate to address the aims of the research?
Item 4: Was the recruitment strategy appropriate to the aims of the research?
Item 5: Was the data collected in a way that addressed the research issue?
Item 6: Has the relationship between researcher and participants been adequately considered?
Item 7: Have ethical issues been taken into consideration?
Item 8: Was the data analysis sufficiently rigorous?
Item 9: Is there a clear statement of findings?
Item 10: How valuable is the research?

The responses to items: Yes, Can't Tell, No
µ Overall quality of study: high", "moderate" or "low" was evaluated by the review team based on how reliable and credible each study was, without specific rules

Supplement 3: Quality of distribution studies assessed using the National Institute of Health (NIH) quality assessment tool for before-after (pre-post) studies with no control group[31]

Study (Author, Year)	Criteria 1 #	Criteria 2 #	Criteria 3 #	Criteria 4 #	Criteria 5 #	Criteria 6 #	Criteria 7 #	Criteria 8 #	Criteria 9 #	Criteria 10 #	Criteria 11 #	Criteria 12 #	Overall quality µ
Alghadir, 2017	Yes	Yes	Yes	Yes	Other	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Alghadir, 2018	Yes	Yes	Yes	Yes	Other	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Alghadir, 2016 b	Yes	No	Yes	Yes	Yes	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Alghadir, 2017	Yes	Yes	Yes	Yes	Other	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Baert, 2018	Yes	No	Yes	Yes	Other	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Baert, 2018	Yes	Yes	Yes	Yes	Other	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Brisson, 2018	Yes	Yes	Yes	Yes	Yes	Other	Yes	Other	Yes	Yes	Yes	Other	**Good**
Callaghan, 2009	Yes	No	Yes	Yes	Other	Other	Yes	Other	Other	Yes	Yes	Other	**Fair**
Hoglund, 2019	Yes	Yes	Yes	Yes	Yes	Other	Yes	Other	Yes	Yes	Yes	Other	**Good**
Hunter, 2006	Yes	Yes	Yes	Yes	Other	Other	Yes	Other	Other	Yes	Yes	Other	**Fair**
Ijima, 2019	Yes	Yes	Yes	Yes	Other	Yes	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Jansen, 2021	Yes	Yes	Yes	Yes	Other	Yes	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Kanko, 2019	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Other	Yes	Yes	Yes	Other	**Good**
Kean, 2010	Yes	Yes	Yes	Yes	Other	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Klokker, 2015	Yes	Yes	Yes	Yes	Yes	Other	Yes	Other	Yes	Yes	Yes	Other	**Good**
McCarthy, 2004	Yes	Yes	Yes	Yes	Other	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
McCarthy, 2008	Yes	Yes	No	No	Other	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Monticone, 2021	Yes	Yes	Yes	Yes	Yes	Other	Yes	Other	Yes	Yes	Yes	Other	**Good**
Motyl, 2013	Yes	Yes	Yes	Yes	Yes	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Mutlu, 2015	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Other	Yes	Yes	Yes	Other	**Good**
Nalbant, 2021	Yes	Yes	Yes	Yes	Yes	Other	Yes	Other	Yes	Yes	Yes	Other	**Good**
Naylor, 2014	Yes	Yes	Yes	Yes	Other	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Parveen, 2017	Yes	Yes	Yes	Yes	Yes	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Peter, 2018	Yes	Yes	Yes	Yes	Other	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Piva, 2004	Yes	Yes	Yes	Yes	Other	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Pratheep, 2018	Yes	Yes	Yes	Yes	Other	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Ravaud, 1999	Yes	Yes	Yes	Yes	Other	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**

Study (Author, Year)	Criteria 1 #	Criteria 2 #	Criteria 3 #	Criteria 4 #	Criteria 5 #	Criteria 6 #	Criteria 7 #	Criteria 8 #	Criteria 9 #	Criteria 10 #	Criteria 11 #	Criteria 12 #	Overall quality µ
Suhail and Chaudhary, 2021	Yes	Yes	Yes	Yes	Yes	Other	Yes	Other	Yes	Yes	Yes	Other	**Good**
Suwit, 2020	Yes	Yes	Yes	Yes	Other	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Takacs, 2014	Yes	Yes	Yes	Yes	Other	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Tevald, 2016	Yes	Yes	Yes	Yes	Other	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Takacs, 2014	Yes	Yes	Yes	Yes	Yes	Other	Yes	Other	Yes	Yes	Yes	Other	**Good**
Tevald, 2016	Yes	No	Yes	Yes	Other	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Tse, 2021	Yes	Yes	Yes	Yes	Other	Other	Yes	Other	Yes	Yes	Yes	Other	**Fair**
Turcot, 2008	Yes	Yes	Yes	Yes	Yes	Other	Yes	Other	Yes	Yes	Yes	Other	**Good**

The description of the criteria is given below.
Criteria 1: Was the study question or objective clearly stated?
Criteria 2: Were eligibility/selection criteria for the study population prespecified and clearly described?
Criteria 3: Were the participants in the study representative of those who would be eligible for the test/service/intervention in the general or clinical population of interest?
Criteria 4: Were all eligible participants that met the prespecified entry criteria enrolled?
Criteria 5: Was the sample size sufficiently large to provide confidence in the findings?
Criteria 6: Was the test/service/intervention clearly described and delivered consistently across the study population?
Criteria 7: Were the outcome measures prespecified, clearly defined, valid, reliable, and assessed consistently across all study participants?
Criteria 8: Were the people assessing the outcomes blinded to the participants' exposures/interventions?
Criteria 9: Was the loss to follow-up after baseline 20% or less? Were those lost to follow-up accounted for in the analysis?
Criteria 10: Did the statistical methods examine changes in outcome measures from before to after the intervention? Were statistical tests done that provided p values for the pre-to-post changes?
Criteria 11: Were outcome measures of interest taken multiple times before the intervention and multiple times after the intervention (i.e., did they use an interrupted time-series design)?
Criteria 12: If the intervention was conducted at a group level (e.g., a whole hospital, a community, etc.) did the statistical analysis take into account the use of individual-level data to determine effects at the group level?

The responses to items: Yes, No, Other (CD, NR, NA)*, *CD, cannot determine; NA, not applicable; NR, not reported
µ Overall quality: Good, Fair, or Poor was evaluated by the review team based on how reliable and credible each study was, without specific rules as recommended by the tool

Supplement 4: Anchor properties of included anchor studies

Study	Description of anchor	Anchor question	Anchor properties/responses	Definitions of MIC/MID using transition question	Reported terminology in the study
Angst, 2018	GRC transition (5 points)	NR	much worse, slightly worse, almost equal, slightly better, much better	Difference between the "slightly better" group and the "almost equal" group= MID	Minimal Clinically Important Difference
Harris, 2013	GRC responses (3 points)	Compared to one week before your clinical visit, please indicate how much your knee problem has changed?	1. My knee has got better, 2. My knee has stayed the same, 3. My knee has got worse	Mean change in group "my knee has got better"= MIC and the difference in change score between groups responded with " my knee has stayed the same" and "my knee has got better"= MID	Minimal Important Change, Minimal Important Difference
Hmamouchi, 2012	GRC transition (5 points)	How do you feel in general today as compared to six weeks earlier as far as your osteoarthritis is compared?	much better, slightly better, no change, slightly worse, much worse	Difference between the mean effects of "slightly better' group and "no change" groups= MID	Minimal Clinically Important Difference for improvement
Klokker, 2016	modified GRC (3 responses and a GRC spanning from -7 to +7)	Did your knee pain change since you entered this project?	unchanged, better, worse: if it is better or worse, bring up a scale spanning from -7 to +7 (-7 (worst) to +7 (best) scale)	Difference of a score of at least 2 (+2: little better: -2: little worse) and no change (score of 0, +1 (almost the same or hardly any better), -1 (worse at all)) = MID	Minimal Important Change
Lee, 2018	Multiple anchors: 36 item Short Form subscales- physical functioning, social role functioning, energy and vitality, bodily pain, mental health, physical role functioning, emotional role functioning, general health perception (0-100	NR	NR	Prospective change for people achieving previously established MID on legacy comparators: MID in a single item anchor e.g. Patient Global Assessment was defined as range= MID to 1+MID, MID in a multi-teem anchor e.g. SF36 range= MID to 2XMID	Minimally Important Difference

Study	Description of anchor	Anchor question	Anchor properties/responses	Definitions of MIC/MID using transition question	Reported terminology in the study
	scale, 0: best, 100: worse); WOMAC-pain and function (0-100 scales,0: best, 100: worse); Back depression (0-63, 0: best, 63: worse); Perceived Stress scale (0-40, 0: best, 40: worse); Patient Global Assessment in a 0-10 cm visual analogue scale (higher number= greater perception of disease activity); Six-minute walk test (in meters); 20-meter walk test (in seconds)				
Mills, 2016	Global transition GRC-7-point Likert scale	Two anchor question 1. Compared with when I started this program, my walking on level ground has (walking anchor) and 2. Compared with when I started this program, my knee had (knee health anchor):	much improved, moderately improved, slightly improved, not changed, slightly worse, moderately worse, much worse	Mean change within slightly improved group= MIC And Difference between participants who responded slightly, moderately, much improved as the "improved group" and no change or worse "non-improved group" =MID	Minimal Important Difference-Mean change (Jaeschke) method Minimal Important Difference-Mean change (Redelmeier and Lorig) method, ROC method (Youden method and 80% specific method)
Mostafee, 2021	GRC- 7points Likert scale	How did your knee status change compared to the beginning of the	1. very much worse; 2. much worse; 3. slightly worse; 4. no change; 5. slightly better; 6. much	The difference between improved group (6 and 7) and not improved group (1,2,3,4 and 5)=MID	Minimal Important Change

Study	Description of anchor	Anchor question	Anchor properties/responses	Definitions of MIC/MID using transition question	Reported terminology in the study
		physiotherapy intervention?	better; and 7. very much better		
Ornetti, 2011	GRC- (3points Likert scale, then again in a 4-point Likert scale	specific question NR. Two anchor questions. 1. The degree of improvement of global state 2. The degree of improvement of functional state	The degree of improvement of global state (global MCII) on a 3-point Likert scale (worsened function, no change, improved function). Then, among the patients who improved, the degree of improvement was scored on a scale (poor, fair, good, excellent)	75th centile of the absolute change in score among patients who responded the improvement as good or excellent (improved group) = MIC	Minimal Clinically Important Improvement
Perrot, 2013	GRC (5 points)	Taking into account the pain due to your arthritis in the last 24 hours and the pain you experienced initially, how has your pain changed?	Patients who said that their pain had improved were asked to rate the level of improvement on a 5-point Likert scale extending from very large improvement to no improvement at all	75th percentile of the distribution of the pain intensity difference (day 0 to 7) for patients considering their improvement to be at least moderate= MIC	Minimal Clinically Important Improvement
Singh, 2014	GRC (5 points)	Since the last time you completed the survey 2 weeks ago, would you say your knee arthritis is:	A great deal better, somewhat better, about the same, somewhat worse, a great deal worse	Mean change between baseline and follow-up of the group who said somewhat better= MIC	Minimal Clinically Important Difference for improvement

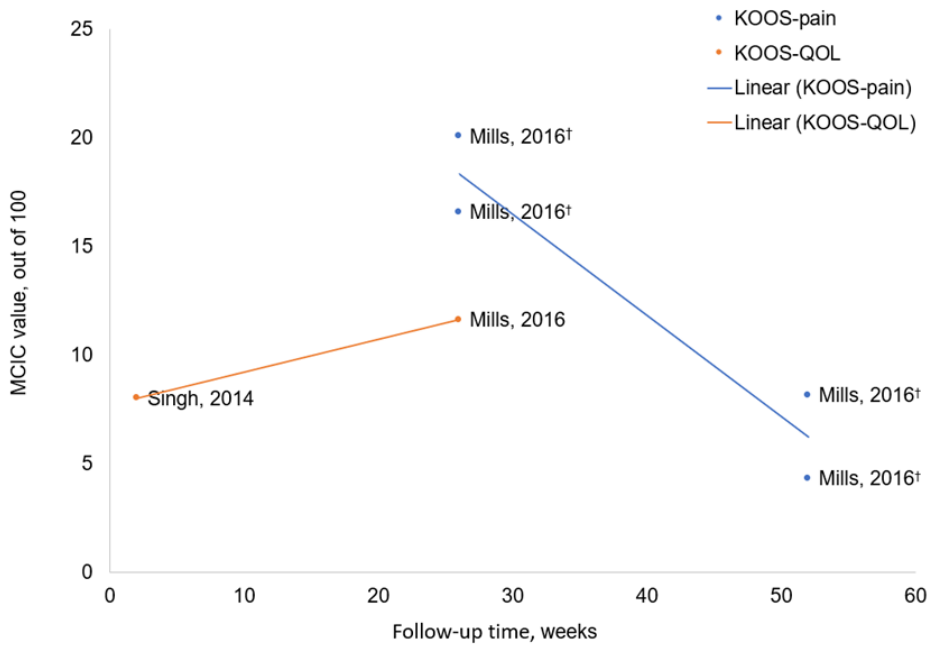
Study	Description of anchor	Anchor question	Anchor properties/responses	Definitions of MIC/MID using transition question	Reported terminology in the study
Tubach, 2005	GRC (5 points)	Response to NSAID treatment	None= no good at all, ineffective drug; poor= some effect but unsatisfactory; fair= reasonable effect but could be better; good= satisfactory effect with occasional episodes of pain or stiffness; excellent= ideal response, virtually pain-free	75th centile of the change in score of the group who reported as “good” using logistic regression=MIC	Minimal Clinically Important Improvement
Williams, 2012	GRC questionnaire on a 15-level scale (15 points)	Not specified To rate the extent to which they perceive their condition as having changed over time	GRC on a 15- level scale. The GRC ranges from 1= a very great deal better to 8=about the same to 15= very great deal worse	Difference between subjects who perceived "improved" (those with a GRC between 1 (very great deal better) and 5 (somewhat better)) from subjects who perceived "not improved" (those with between 6 (a little better) and 15 a very great deal worse)) =MID	Minimal Clinically Important Difference

MIC: Minimal Important Change, MID: Minimal Important Difference, GRC: Global Rating of Change, NR: Not Reported, CI: Confidence Interval, WOMAC: Western Ontario and McMaster Universities Arthritis Index, SF36: Short Form-36, NSAID: Non-Steroid Anti-Inflammatory Drug

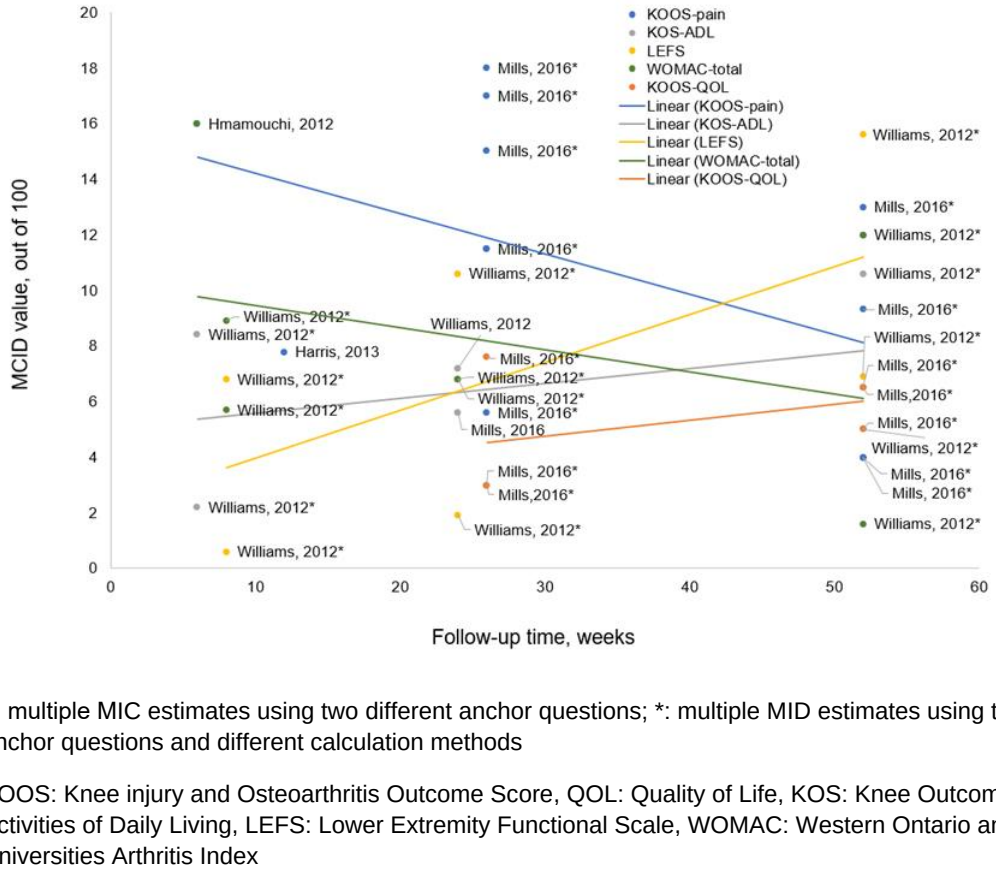
10.1136/bmjopen-2023-073320 on 18 May 2023. Downloaded from

Supplement 5: The effect of follow-up time on **A.** Minimal Important Change (MIC) and **B.** Minimal Important Difference (MID)

A.



B.



†: multiple MIC estimates using two different anchor questions; *: multiple MID estimates using two different anchor questions and different calculation methods

KOOS: Knee injury and Osteoarthritis Outcome Score, QOL: Quality of Life, KOS: Knee Outcome Survey, ADL: Activities of Daily Living, LEFS: Lower Extremity Functional Scale, WOMAC: Western Ontario and McMaster Universities Arthritis Index

Supplement 6: Minimum Detectable Change (MDC) values of knee osteoarthritis outcome tools derived using the distribution method

Outcome tool	Score	Median	Minimum	Maximum	Number of estimates	Study
Aggregated locomotor function	N/A	2.3 seconds			1	McCarthy,2004
Animated Activity Questionnaire	100=best	11.2			1	Peter, 2018
BMI using a scale and a stadiometer	N/A	2.6 kg/m^2			1	Brisson, 2018
Bone density-femur mean lateral	N/A	0.5 mm	0.4 mm	0.6 mm	2	Jansen, 2021
Bone density-femur mean medial	N/A	0.6 mm Al eq	0.5 mm Al eq	0.7 mm Al eq	2	Jansen, 2021
Bone density-tibia mean lateral	N/A	2.6 mm Al eq	2.4 mm Al eq	2.8 mm Al eq	2	Jansen, 2021
Bone density-tibia mean medial	N/A	0.8 mm Al eq	0.7 mm Al eq	1 mm Al eq	2	Jansen, 2021
Cartilage volume-lateral tibia	N/A	0.6ml	0.5 ml	0.6 ml	2	Hunter, 2006
Cartilage volume-medial Tibia	N/A	0.6 ml	0.4 ml	0.7 ml	2	Hunter, 2006
Cartilage volume-femur	N/A	1.1 ml	0.8 ml	1.3 ml	2	Hunter, 2006
Cartilage volume-patella	N/A	0.9 ml	0.7 ml	1.1 ml	2	Hunter, 2006
Center of mass mediolateral displacement during unipodal stance using 3D motion analysis	N/A	0.01º	0.01º	0.02 º	3	VandeStraaten, 2020
De Motion Mobility Index	100=best	8.0	7.3	8.7	2	Yuruk, 2014
Eight-meter walk time	N/A	1.4 seconds			1	McCarthy,2004
Eminence lateral (knee image)	mm	2.2 mm	2.2 mm	2.2 mm	2	Jansen, 2021
Eminence medial (knee image)	mm	1.8 mm	1.7 mm	1.8 mm	2	Jansen, 2021
Get up and Go test	N/A	1.4 seconds	1.2 seconds	1.5 seconds	2	Piva, 2004
Fremantle Knee Awareness Questionnaire	100=worst	14.4			1	Monticone,2021
Frontal plane tibial alignment using smartphone inclinometer	N/A	3.7º			1	Tse, 2021
Frontal plane tibial alignment using a manual inclinometer	N/A	3.2º			1	Tse, 2021
Hip abductor strength using a hand-held dynameter	N/A	0.3 Nm/kg	0.3 Nm/kg	0.3 Nm/kg	2	Tevald, 2016
Hip flexion-extension during unipodal stance using 3D motion analysis	N/A	2.7 º	2.2 º	4.6 º	3	VandeStraaten, 2020
ICOAP-constant pain	100=worst	51.7	49.6	53.8	2	Singh, 2014
ICOAP-intermittent pain	100=worst	49.8	48.7	50.8	2	Singh, 2014
ICOAP-pain	100=worst	46.6	23.0	49.6	3	Singh, 2014, Harris, 2013
KAM impulse in 3D motion analysis	N/A	4.9 Nm*s			1	Brisson, 2018
KAM impulse in 3D motion analysis	N/A	0.4 %BW*HT*s			1	Brisson, 2018
Knee force sense test using a handheld dynamometer-reposition error 20º	N/A	20.6 N	13.2 N	26.7 N	3	Baert, 2018

Outcome tool	Score	Median	Minimum	Maximum	Number of estimates	Study
Knee force sense test using a handheld dynamometer-reposition error 45º	N/A	38.0 N	27.8 N	43.3 N	3	Baert, 2018
Knee force sense test using a handheld dynamometer-reposition error 70º	N/A	32.2 N	22.5 N	49.5 N	3	Baert, 2018
Knee force sense test using a handheld dynamometer-reposition error all	N/A	32.7 N	30.0 N	33.7 N	3	Baert, 2018
Knee joint angle	degrees	2.1º	1.9º	2.3º	2	Jansen, 2021
Knee joint space width	mm	1 mm	0.8 mm	1.2 mm	2	Jansen, 2021
Knee joint space width-lateral	mm	1.7 mm	1.4 mm	2.0 mm	2	Jansen, 2021
Knee joint space width-medial	mm	0.6 mm	0.5 mm	0.8 mm	2	Jansen, 2021
Knee joint space width-minimum	mm	0.8 mm	0.6 mm	0.9mm	2	Jansen, 2021
KOOS-activities of daily living	100=best	19.1	17.4	20.8	2	Naylor, 2014
KOOS-pain	100=best	18.6	17	20.2	2	Naylor, 2014
KOOS-quality of life	100=best	27.8	22.4	39.0	4	Naylor, 2014; Singh, 2014
KOOS-symptoms	100=best	20.2	2.9	24.1	3	Naylor, 2014
KOOS-PS	100=worst	28.3	16.0	35.5	3	Harris, 2013 Singh, 2014
Knee joint position sense test-analogue inclinometer-reposition error 70º	N/A	8.0 º	4.0 º	8.0 º	3	Baert, 2018
Knee joint position sense test-analogue inclinometer-reposition error 45º	N/A	4.0 º	3.0 º	8.0 º	3	Baert, 2018
Knee joint position sense test-analogue inclinometer-reposition error 20º	N/A	4.0 º	3.0 º	4.0 º	3	Baert, 2018
Knee joint position sense test-analogue inclinometer-reposition error all	N/A	3.0 º	3.0 º	4.0 º	3	Baert, 2018
KOS-activities of daily living	100=best	15.8*	10.5	21.0	6	Williams, 2012
L-test	N/A	5.28 seconds			1	Nalbant, 2021
Load frequency using a triaxial accelerometer	N/A	4.3 steps/day			1	Brisson, 2018
LEFS	100=best	18.3*	14.8	22.6	6	Williams, 2012
NRS- pain	100=worst	16.5	13.3	19.6	2	Alghadir, 2016b, Alghadir, 2018
Osteophytes-Femur lateral (using knee image)	N/A	7.8 mm^2	5.4 mm^2	10.3 mm^2	2	Jansen, 2021
Osteophytes-Femur medial (using knee image)	N/A	12.3 mm^2	11.2 mm^2	13.4 mm^2	2	Jansen, 2021
Osteophytes-Tibia lateral (using knee image)	N/A	10.5 mm^2	9.4 mm^2	11.6 mm^2	2	Jansen, 2021

Outcome tool	Score	Median	Minimum	Maximum	Number of estimates	Study
Osteophytes-Tibia medial (using knee image)	N/A	11.6 mm^2	11 mm^2	12.2 mm^2	2	Jansen, 2021
OKS-summary	100=best	6.1	6.0	6.2	1	Alghadir, 2017, Harris, 2013
Peak KAM in 3D motion analysis	N/A	0.2 Nm/kg			1	Brisson, 2018
Peak KAM in 3D motion analysis	N/A	1.3 %BW*HT			1	Brisson, 2018
Peak KFM in 3Dmotion analysis	N/A	0.4 Nm/kg			1	Brisson, 2018
Peak KFM in 3D motion analysis	N/A	2.3 %BW*HT			1	Brisson, 2018
Pelvic abduction-adduction during unipodal stance using 3D motion analysis	N/A	3.1 º	2.8 º	3.5 º	3	VandeStraaten, 2020
Per cent of voluntary muscle activation using a dynamometer	N/A	6.6%			1	Kean, 2010
Pressure pain threshold-knee- using an algometer	N/A	131.8 kPa	92.9 kPa	196.3 kPa	10	Pratheep, 2018
Pressure pain threshold-medial heel using a handheld pressure algometer	N/A	1.3 lb			1	Mutlu, 2015
Pressure pain threshold- medial knee using a handheld pressure algometer	N/A	1.2 lb			1	Mutlu, 2015
Quadriceps fatigue using surface electromyography-VMO initial median frequency	N/A	11.0 Hz			1	Callghan, 2009
Quadriceps fatigue using surface electromyography-VL initial median frequency	N/A	10.0 Hz			1	Callghan, 2009
Quadriceps fatigue using surface electromyography-RF initial median frequency	N/A	9.0 Hz			1	Callghan, 2009
Quadriceps fatigue using surface electromyography-VMO final median frequency	N/A	7.4 Hz			1	Callghan, 2009
Quadriceps fatigue using surface electromyography-VL final median frequency	N/A	6.5 Hz			1	Callghan, 2009
Quadriceps fatigue using surface electromyography-RF final median frequency	N/A	10.5 Hz			1	Callghan, 2009
Quadriceps fatigue using surface electromyography-VMO median frequency slope	N/A	2207.0 %/min*			1	Callghan, 2009
Quadriceps fatigue using surface electromyography-VL median frequency slope	N/A	4000.0 %/min*			1	Callghan, 2009
Quadriceps fatigue using surface electromyography-RF median frequency slope	N/A	2390.0 %/min*			1	Callghan, 2009

Outcome tool	Score	Median	Minimum	Maximum	Number of estimates	Study
Quadriceps isokinetic strength using a dynamometer-absolute value	N/A	33.9 Nm			1	Kean, 2010
Quadriceps isometric strength using a dynamometer-absolute value	N/A	25.0 Nm	5.5	37.2	3	McCarthy,2008; Brisson, 2018
Quadriceps isometric strength using a dynamometer-normalised to weight	N/A	0.4 Nm/kg	0.3	0.5	2	Brisson, 2018; Kean, 2010
Quadriceps isometric strength using a dynamometer-Normalised to body size	N/A	1.5 %BW*height			1	Kean, 2010
Quadriceps power using a dynamometer	N/A	151.8 W/2.2 W/kg			1	Brisson, 2018
Rectus femoris fatigue slope	N/A	0.6 %/min			1	McCarthy,2008
Rectus femoris initial median frequency	N/A	5.2 Hz			1	McCarthy,2008
Single-leg standing balance test-mediolateral standard deviation	N/A	0.3			1	Takacs, 2014
Single-leg standing balance test-anteroposterior standard deviation	N/A	0.5			1	Takacs, 2014
Single-leg standing balance test-path length	N/A	20.2 cm			1	Takacs, 2014
Single-leg standing balance test-velocity	N/A	0.3 m/seconds			1	Takacs, 2014
Single-leg standing balance test-area	N/A	23.2 cm^2			1	Takacs, 2014
Six minute walk test	N/A	72.7 m	66.3 m	79.0 m	2	Naylor, 2014
Stair ascent and descent time (seven steps (four of 15cm, three of 20cm))	N/A	2.6 seconds			1	McCarthy,2004
Stopwatch-based 11- Step stair climb test	N/A	0.1 seconds	0.1 seconds	0.1 seconds	2	Ijima, 2019
Star excursion balance test-Raw value	N/A	7.4 cm			1	Kanko, 2019
Star excursion balance test-Normalized (raw value/leg length%)	N/A	8.5 %			1	Kanko, 2019
Timed Up and Go test	N/A	1.1 seconds	1.1 seconds	1.1 seconds	2	Alghadir, 2015
Timed Up and Go test (as a ratio of the original measurement)	N/A	41.06%	37.5%	44.6%	2	Naylor, 2014
Tinetti Performance-Oriented Mobility Assessment scale- balance subscale	NR	0.8			1	Parveen, 2017
Tinetti Performance-Oriented Mobility Assessment scale- gait subscale	NR	0.6			1	Parveen, 2017
Tinetti Performance-Oriented Mobility Assessment scale- total	NR	1.0			1	Parveen, 2017

Outcome tool	Score	Median	Minimum	Maximum	Number of estimates	Study
Transferring time (distance of 2m to a chair and sit down, then, walk back to the start)	N/A	1.7			1	McCarthy,2004
Trunk-abduction-adduction during unipodal stance using 3D motion analysis	N/A	2.9 º	2.6 º	2.9 º	3	VandeStraaten, 2020
2-minute walk test	N/A	5.52m			1	Suhail and Chaudhary, 2021
Vastus lateralis fatigue slope	N/A	0.5 %/min			1	McCarthy,2008
Vastus lateralis initial median frequency	N/A	5.8 Hz			1	McCarthy,2008
Vastus medialis oblique fatigue slope	N/A	0.8 %/min			1	McCarthy,2008
Vastus medialis oblique initial median frequency	N/A	6.7 Hz			1	McCarthy,2008
Verbal Rating Scale for pain		5.8			1	Alghadir, 2018
VAS for pain	100=worst	24.0 cm *	8.0 cm	28.0 cm	3	Alghadir, 2018; Naylor,2014
WOMAC-pain	100=best	3.8	3.4	19.0	3	Angst, 2018, Alghadir, 2016 a
WOMAC-function	100=best	3.1	3.1	18.7	3	Angst, 2018, Alghadir, 2016 a
WOMAC-stiffness	100=best	5.2	4.8	5.6	2	Angst, 2018
WOMAC-functional standing/walking	100=best	3.8	3.5	4.0	2	Angst, 2018
WOMAC-total	100=worst	18.7	11.7	20.8	6	Williams, 2012, Alghadir, 2016 a
SF-36-bodily pain	100=best	3.5	3.3	3.7	2	Angst, 2018
SF-36-general health	100=best	2.1	2.0	2.2	2	Angst, 2018
SF-36-mental health	100=best	3.1	2.9	3.4	2	Angst, 2018
SF-36-physical functioning	100=best	2.4	2.3	2.4	2	Angst, 2018
SF-36-role- physical	100=best	11.2	10.9	11.5	2	Angst, 2018
SF-36-social functioning	100=best	4.1	3.7	4.5	2	Angst, 2018
SF-36-vitality	100=best	5.3	5.1	5.6	2	Angst, 2018
20-meter walk test	N/A	0.9 seconds	0.2 seconds	1.6 seconds	2	Motyl, 2013
30-second fast-paced walk test	N/A	3.2 m	0.8 m	5.7 m	2	Hoglund, 2019
40-meter fast-paced test	N/A	16.3			1	Suwit, 2020
30 seconds chair stand test	N/A	21.2			1	Suwit, 2020

Outcome tool	Score	Median	Minimum	Maximum	Number of estimates	Study
3D linear accelerations of the tibia during comfortable walking	N/A	0.3 g	0.1 g	0.8 g	12	Turcot, 2008
3D linear accelerations of the femur during comfortable walking	N/A	0.3 g	0.1 g	1.0 g	12	Turcot, 2008
3D linear accelerations of the tibia at fast speed	N/A	0.4 g	0.1 g	0.9 g	12	Turcot, 2008
3D linear accelerations of the femur at fast speed	N/A	0.2 g	0.0 g	0.6 g	12	Turcot, 2008
9-step stair climb test	N/A					Suwit, 2020

MDC: Minimum Detectable Change, OA: osteoarthritis, BMI: Body Mass Index, MRI: Magnetic Resonance Imaging, 3D: 3-dimensional, N/A: Not Applicable, ICOAP: Intermittent and constant osteoarthritis pain, KOOS: Knee injury and Osteoarthritis Outcome Score, KOOS-PS: Knee injury and Osteoarthritis Outcome Score Physical Function Short form, KOS: Knee Outcome Survey, LEFS: Lower Extremity Functional Scale, NRS: Numeric Rating Scale, OKS: Oxford Knee Score, KAM: Knee adduction moment, KFM: Knee Flexion Moment, VMO: Vastus Medialis Oblique, VL: Vastus Lateralis, RF: Rectus Femoris, VAS: Visual Analog Scale, WOMAC: Western Ontario and McMaster Universities Arthritis Index, SF-36: 36-item Short Form health survey,

All the estimates are out of 100.

The median estimates are shaded in blue

PRISMA 2020 Checklist

Section and Topic	Item #	Checklist item	Location where item is reported
TITLE			
Title	1	Identify the report as a systematic review.	Page 2
ABSTRACT			
Abstract	2	See the PRISMA 2020 for Abstracts checklist.	Page 2 to 3
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	Page 5 to 7
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	Page 7, line 103 to 107
METHODS			
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	Page 8 to 9, line 133 to 153
Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	Page 7, line 116 to 121
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	Page 7 to 8, line 121 to 131
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	Page 8, line 125 to 131
Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	Page 8 -9, line 133 to 153;
Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	Page 10, line 173 to 182
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	Page 10, line 173 to 176
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	Page 9 to 10, line 163 to 183
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	Page 10, line 184 to 188
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	Page 10, line 184 to 188
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	Page 10 to 11, line 184 to 196
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	Page 10 to11, line 184 to 196
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	Page 10, line 184 to 188

PRISMA 2020 Checklist

Section and Topic	Item #	Checklist item	Location where item is reported
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	Not done
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	Not applicable
Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	Not applicable
Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	Not applicable
RESULTS			
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.	Page 11, line 203 to 207 and figure 1
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	Figure 1
Study characteristics	17	Cite each included study and present its characteristics.	Page 11 to 15, line 218 to 256 and table 1
Risk of bias in studies	18	Present assessments of risk of bias for each included study.	Page 11, line 212 to 216 and supplements1,2 and 3
Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	Tables 2,3 and 4
Results of syntheses	20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies.	Page 14 to 18, and table 1, supplements1,2 and 3
	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	Page 18 to 24; tables 2,3,4 and supplements 5,6
	20c	Present results of all investigations of possible causes of heterogeneity among study results.	Page 18 to 24;, tables 2,3,4 and supplements 5,6
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	Not applicable
Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	Not applicable
Certainty of evidence	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	Not applicable
DISCUSSION			

PRISMA 2020 Checklist

Section and Topic	Item #	Checklist item	Location where item is reported
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.	Page 26 to 30, line 399 to 503
	23b	Discuss any limitations of the evidence included in the review.	Pages 29 to 30, line 483 to 491
	23c	Discuss any limitations of the review processes used.	Page 29 to 30, line 491 to 494
	23d	Discuss implications of the results for practice, policy, and future research.	Page 30, line 496 to 503
OTHER INFORMATION			
Registration and protocol	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	Page 7, line 110 to 114
	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	Page 7, line 110 to 114
	24c	Describe and explain any amendments to information provided at registration or in the protocol.	None
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	None
Competing interests	26	Declare any competing interests of review authors.	None
Availability of data, code and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	None

From: Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. doi: 10.1136/bmj.n71

For more information, visit: http://www.prisma-statement.org/