Thesis Title: Accuracy and Reliability of Remote Shoulder Motion Capturing Methods: A Systematic Review and Meta-Analysis

Volume: 1 of 1

Name: Pengchi Chen

Qualification thesis submission for: MSc by Thesis

The University of Hull & the University of York

Hull York Medical School

August 2025

ABSTRACT

Background:

The COVID-19 pandemic accelerated the demand for remote assessment tools in rehabilitation, spotlighting the need for accurate and reliable technologies to measure shoulder range of motion (ROM) outside of clinical environments. Emerging tools such as smartphone apps, wearable sensors, and markerless motion capture systems are increasingly being adopted, yet their accuracy and reliability compared to reference standards remains unclear.

Objective:

To systematically evaluate the accuracy and reliability of existing remote shoulder ROM measurement technologies, quantify measurement bias, and assess their agreement with reference standards.

Methods:

A systematic review and meta-analysis was conducted on 26 studies evaluating remote ROM measurement tools. Pooled mean bias (in degrees) was calculated as the primary effect size for agreement, with reliability assessed using intraclass correlation coefficients (ICCs). Subgroup analyses were performed by motion type, technology category, population health status, and data acquisition method. Risk of bias was assessed using the QUADAS-2 tool.

Results:

Remote measurement methods showed a small but consistent overestimation of ROM compared to reference standards (pooled mean bias = 2.63°, 95% CI: 1.52°, 3.74°), particularly in flexion, internal rotation, and external rotation. No significant bias was observed in abduction or extension. Both inertial measurement unit (IMU) and non-IMU technologies demonstrated comparable levels of overestimation. Pathological populations exhibited greater variability (bias = 4.33° vs. 2.37° in healthy subjects). Self-measurements showed lower and non-significant bias compared to assessor-guided methods. Reliability was generally high, especially for test-retest assessments (ICCs > 0.90), though more variable in inter-rater and pathological settings.

Conclusion:

Remote technologies for assessing shoulder range of motion show generally small differences compared to reference standards, but performance is inconsistent, with substantial heterogeneity and variable reliability across studies. Although average error is often below reported minimal clinically important differences (MCID), the wide variability observed suggests these tools may not reliably detect small but clinically meaningful changes. As such, they may have value for tracking overall trends in shoulder function, but their use in clinical decision-making—particularly in pathological populations—should be approached with caution.

Table of Contents

List of Tables and Figures

Tables

Figures

Acknowledgement & Declaration

First and foremost, I would like to express my deepest gratitude to my supervisors, Professor Amar Rangan and Dr. Peter Ellison, for their invaluable guidance and expertise throughout the course of this project. Their mentorship has been instrumental in shaping both this thesis and my development as a clinical academic.

I would also like to thank my colleague Miss Mia Prosser for her vital contributions as second reviewer during the systematic review process. Her attention to detail and thoughtful input were essential in strengthening the methodological rigour of the work.

My sincere thanks to Professor Bob Phillips, who served as my TAP Chair, for his ongoing support, critical feedback, and mentorship throughout my academic training.

I am also grateful to Dr. Emily Shoesmith, Dr. Elena Ratschen, Professor Mona Kanaan, and Miss Ada Keding at the University of York for their teachings across key modules including evidence synthesis, regression analysis, and measurement in health and disease. Their instruction provided me with a solid methodological and statistical foundation, not only for this thesis but for my future research endeavours.

I am grateful to the NIHR Academic Clinical Fellowship Programme, which provided the opportunity and resources for me to undertake this research alongside my clinical training.

Lastly, I would like to thank my family and friends for their unwavering support and encouragement during the highs and lows of this academic endeavour.

I confirm that this work is original and that if any passage(s) or diagram(s) have been copied from academic papers, books, the internet or any other sources these are clearly identified by the use of quotation marks and the reference(s) is fully cited. I certify that, other than where indicated, this is my own work and does not breach the regulations of HYMS, the University of Hull or the University of York regarding plagiarism or academic conduct in examinations. I have read the HYMS Code of

Practice on Academic Misconduct, and state that this piece of work is my own and does not contain any unacknowledged work from any other sources.

Introduction

Shoulder range of motion (ROM) assessment is a fundamental component of musculoskeletal and orthopaedic evaluations. [44] Accurate measurement of ROM is essential for diagnosing shoulder conditions, monitoring treatment outcomes, and guiding rehabilitation. Traditional shoulder ROM assessments are typically performed in clinical settings using tools such as handheld goniometers or inclinometers, requiring direct patient-clinician interaction.

A handheld goniometer is a mechanical device that measures joint angles by aligning its arms with anatomical landmarks, typically requiring manual positioning by a clinician. In contrast, inclinometers measure angular displacement relative to gravity, allowing for more objective quantification of joint movement, particularly in single-plane motions. [26] More recently, digital and remote measurement technologies have been developed. Smartphone-based applications often utilise built-in accelerometers and gyroscopes to function as digital inclinometers or employ camera-based pose estimation algorithms to derive joint angles from images or video. [11, 35] Wearable inertial measurement units (IMUs) incorporate accelerometers, gyroscopes, and sometimes magnetometers to track segment orientation and movement in three-dimensional space. [7, 38] Markerless motion capture systems use computer vision techniques to estimate body joint positions without the need for physical markers, typically through depth sensors or standard RGB cameras. [7, 53] These technologies differ in their underlying measurement principles, required user input, and susceptibility to sources of error such as sensor drift, soft tissue artefact, and camera positioning, which are important considerations when comparing their accuracy and reliability to conventional clinical tools.

With the increasing interest in telerehabilitation and remote monitoring, there is a growing body of research assessing the accuracy and reliability of these technologies compared to traditional face-to-face ROM assessments. Accuracy refers to the accuracy of these tools in measuring ROM compared to reference standard, whereas reliability assesses their consistency across different conditions, including test-retest reliability and inter-rater reliability. [49] Despite the desire to adopt

remote ROM measurement tools in routine clinical practice and research, a comprehensive synthesis of the evidence on their accuracy and reliability is lacking.

The use of digital technologies such as smartphone goniometer apps and wearable devices in ROM assessment has been widely explored, particularly before the COVID-19 pandemic. Werner and colleagues (2014) validated a smartphone clinometer application for measuring shoulder ROM. [54] While Mitchell and colleagues (2014) and Johnson and colleagues (2015) assessed the reliability and validity of smartphone-based goniometry in clinical settings. [23, 28] Rigoni and colleagues (2019) evaluated a wireless inertial motion capture device for ROM assessment. [37] Hayes and colleagues (2015) examined traditional methods such as estimating joint angles from patient photographs. [18] These validation studies generally reported measurement errors within approximately 5°–10° of reference standards and intraclass correlation coefficients (ICCs) often exceeding 0.80, which are commonly considered indicative of good reliability in musculoskeletal assessment. [49] In clinical practice, acceptable measurement error for shoulder ROM is often interpreted in relation to the minimum clinically important difference (MCID), which has been reported to range between approximately 11° and 24° depending on the context and method of assessment. [29]

Therefore, while these early studies suggested that digital measurement tools could achieve clinically acceptable levels of accuracy and reliability under controlled conditions, they were largely conducted in face-to-face or supervised environments. As such, the extent to which these performance benchmarks translate to remote or unsupervised settings remains uncertain. Challenges persist in tele-rehabilitation settings, including low sensitivities, poor inter-rater reliability, and variability in clinical testing accuracy. [5, 15] These limitations highlight the need for further research to optimise assessment techniques.

The broader implications of remote rehabilitation have also been explored. Kane and colleagues (2020) conducted a randomised controlled trial assessing post-operative tele-follow-ups for rotator cuff repair patients, reporting comparable patient satisfaction scores to in-person follow-ups; it did not, however, evaluate longitudinal ROM outcomes, leaving uncertainties regarding the effectiveness of tele-

rehabilitation in monitoring recovery progress. [24] Similarly, Faber and colleagues (2015) highlighted challenges in home-based rehabilitation, particularly in patients with frozen shoulders, who often struggle to follow rehabilitation instructions. These findings reinforce the need for structured and engaging tele-rehabilitation strategies to enhance patient compliance. [13] However, while poor adherence to rehabilitation programmes is well recognised, the role of objective monitoring in addressing this challenge remains less clearly defined. Reliable remote measurement of shoulder range of motion (ROM) may provide clinicians with the ability to track patient progress more accurately, identify suboptimal engagement or incorrect exercise performance, and offer timely feedback or intervention. In this context, the availability of accurate and reliable remote ROM assessment tools may support improved patient engagement by enabling personalised feedback and accountability, thereby complementing rehabilitation strategies rather than replacing them. This highlights the importance of establishing whether such remote measurement technologies can provide clinically acceptable levels of accuracy and reliability, which is the focus of the present review.

Validated patient-reported outcome measures (PROMs) such as the Oxford Shoulder Score (OSS), a 12-item joint-specific questionnaire, are widely utilised in randomised controlled trials assessing treatment outcomes across populations with various shoulder conditions. [32, 36] While PROMs focus on pragmatic aspects of patient outcomes, emphasising pain levels and the ability to perform daily activities as indicators of improved shoulder function, there are concerns that they do not fully reflect the range of motion achieved. [23] Following recommendation by the European Society for Surgery of the Shoulder and Elbow, the Constant-Murley score has been widely used, as it combines subjective patient-reported outcomes and objective range of motion measurements, which contribute to 40% of the total score. [9] While range of motion is a primary outcome in shoulder assessment, it does not fully capture functional movement quality. Muscle strength is an integral component of shoulder function and is often assessed alongside range of motion in both clinical and remote settings. Therefore, studies assessing strength were included where they contributed to the evaluation of shoulder movement performance.

Digital app-based technologies have demonstrated potential in increasing patient engagement following total joint arthroplasty, particularly for hip and knee replacements; their application in shoulder rehabilitation, however, remains unvalidated, representing a gap in the literature. [25] Advanced motion-detecting technologies, such as depth-based skeleton tracking, have shown promise in remote rehabilitation applications but remain impractical for routine clinical use due to technological constraints. [1] Similarly, the accuracy of telehealth shoulder examinations has been critically assessed, with findings suggesting the need for improved methodologies and standardisation in remote assessments. [5]

In the context of this review, remote ROM measurement is defined as assessments conducted without direct physical interaction between the patient and a healthcare provider. This includes self-measurements performed by patients and clinician-guided assessments conducted remotely via digital or telemedicine platforms. This definition distinguishes the present review from previous work, such as systematic reviews assessing the validity of digital ROM measurement devices in clinical settings. [45] This review specifically investigates technologies that facilitate ROM assessments remotely, with minimal or no physical contact between the patient and clinician. This distinction is crucial as it frames the applicability of the findings toward tele-rehabilitation, remote monitoring, and digital health integration rather than merely in-clinic digital measurement advancements.

Objectives

The reserach objectives are summarised using a PICO-style framework:

Population (P): adults with healthy shoulders or musculoskeletal shoulder conditions undergoing shoulder range of motion & strength assessment.
Intervention/Index test (I): remote shoulder ROM & strength measurement methods, including smartphone applications, wearable sensors, and markerless motion capture systems.
Comparator/Reference standard (C): conventional or reference assessment methods, including goniometers, inclinometers, optical motion capture systems, dynamometers and in-person clinician assessment where applicable.

Outcomes (O): accuracy, expressed as measurement bias or agreement with the reference standard, and reliability, including test-retest reliability/intra-rater reliability, inter-rater reliability and agreement among instruments.

Accordingly, this review aims to address the following questions:

1. What is the accuracy of remote shoulder ROM measurement methods compared with reference standard assessments?
2. What is the reliability of remote shoulder ROM measurement tools, including test-retest and inter-rater reliability where reported?
3. How does measurement performance vary according to motion type, technology type, population, and data acquisition methods?

The wording of the thesis objectives has been refined from the original PROSPERO registration to improve clarity and alignment with the final included evidence. These revisions do not change the underlying review question, eligibility criteria, or core outcomes, but rather provide clearer framing of the same review aims for the reader.

Methods

Protocol and Registration

This systematic review follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. The protocol was registered on PROSPERO (Registration Number: CRD42024564283). PRISMA diagram is available as **Figure 1**.

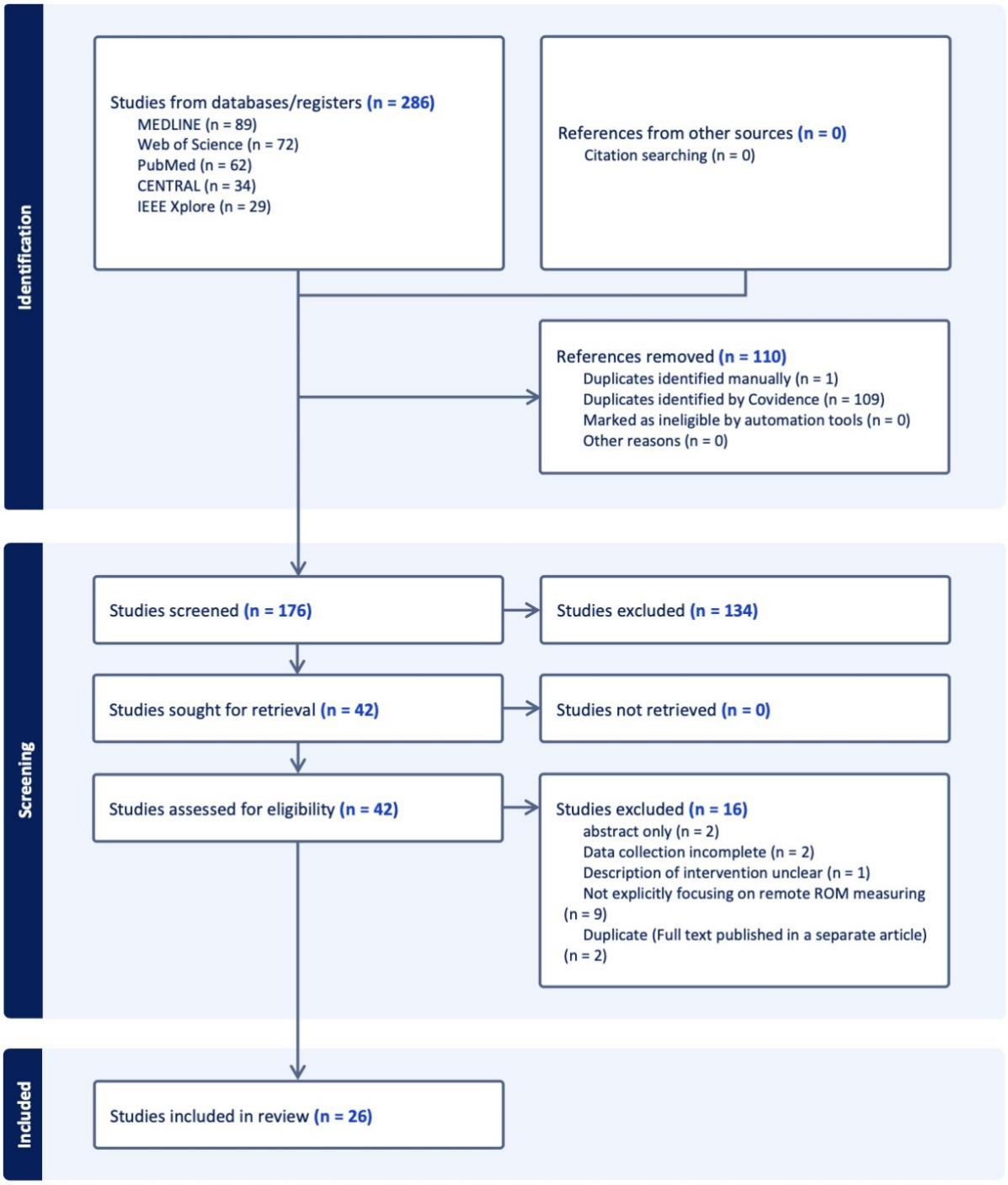

Figure 1 . PRISMA Diagram for study selection

Eligibility Criteria

The inclusion and exclusion criteria for this review were developed collaboratively by the authors, guided by the PICOS framework. Eligible studies focused on adults (18 years or older) diagnosed with musculoskeletal shoulder conditions or healthy volunteers. Remote testing for shoulder ROM and strength using tools such as

wearable devices or mobile applications are included. While control comparators were not required, any that were present in the studies were documented. Key outcomes included accuracy and reliability outcomes reported as mean bias (in degrees) or intraclass correlation coefficient (ICC). Accepted study designs included randomised controlled trials (RCTs), quasi-experimental designs, prospective cohort studies, comparative diagnostic accuracy studies, and case series.

Strength was considered a complementary outcome to range of motion, reflecting overall movement quality and functional capacity of the shoulder. Studies assessing shoulder function, including range of motion and/or muscle strength, were included where these outcomes were relevant to remote assessment of movement.

Only full-text articles published in English were included. Conference abstracts, editorials, and non-peer-reviewed sources were excluded due to limited methodological detail. No restriction was placed on publication date, and databases were initially searched from inception to May 2024.

Participants included both healthy individuals and patients with musculoskeletal shoulder conditions, such as rotator cuff pathology, adhesive capsulitis, post-operative rehabilitation cases, and other shoulder disorders affecting range of motion.

Accuracy outcomes were defined as the agreement between remote measurement methods and reference standards, typically expressed as mean bias or mean difference in degrees, with associated measures of variability such as standard deviation (SD) or limits of agreement (LOA). Reliability outcomes were assessed using intraclass correlation coefficients (ICCs) to evaluate consistency of measurements, including test-retest and inter-rater reliability where reported. The choice of outcome measure was therefore dependent on whether studies assessed agreement with a reference standard (accuracy) or consistency across repeated or multiple measurements (reliability).

Studies were excluded if they assessed motion or strength in non-human subjects (e.g., robotic arms), did not focus explicitly on remote shoulder ROM measurement

methods, or were unrelated to orthopaedic or musculoskeletal conditions. Additionally, studies were excluded if the intervention lacked clarity, or if they were not original research (such as systematic reviews, theses, or dissertations).

Database and Search Strategy

A systematic search was conducted across MEDLINE, Web of Science, PubMed, CENTRAL and IEEE Xplore. The search strategy included terms related to shoulder ROM, remote measurement, accuracy, and reliability. The reference lists of relevant articles were manually screened for additional studies. An example database search using keywords connected by Boolean operators is included as **Appendix 1**. Appendix 1 provides the full search strategies used for each database, including database-specific adaptations of search terms and Boolean operators, to allow replication of the search process.

All databases were initially searched from inception to May 2024. To ensure the currency of the review, an updated search was conducted in March 2026 using the same search strategy. 33 additional studies were identified from all databases after duplicate removal and assessed for eligibility. No additional studies meeting the inclusion criteria were identified and there is no impact on the overall conclusion.

Trial registers (including ClinicalTrials.gov and the WHO International Clinical Trials Registry Platform) were not systematically searched, as the focus of this review was on published studies reporting validated accuracy and reliability outcomes. However, reference list screening was undertaken to minimise the risk of missing relevant studies.

Study Selection

Two independent reviewers (PC & MP) screened titles and abstracts, followed by full-text review. Discrepancies were resolved through discussion or by consulting a third reviewer (PE/AR).

Data Extraction and Risk of Bias Assessment

Data extraction included study characteristics, population details, measurement methods, accuracy and reliability outcomes, and risk of bias assessment. The QUADAS-2 was applied to all studies. [55]

Statistical Analysis

For the quantitative synthesis, meta-analyses were conducted using STATA 18.0 (STATACorp. 2023, TX). Mean bias was selected as the effect size to represent the agreement between remote measurement methods and reference standards. The mean bias is presented as remote methods minus the reference standards. Where applicable, studies reporting mean difference (bias) and standard deviation (SD) were directly included.

For studies that only reported 95% confidence intervals (CI) of the mean bias, the standard deviation was recalculated using the formula:

$$SD = \frac{(Upper\ CI - Lower\ CI)}{3.92}$$

This approach assumes a normal distribution and aligns with established methods for deriving SD from 95% CI ranges. The standard error (SE) was subsequently calculated as:

$$SE = \frac{SD}{\sqrt{n}}$$

where *n* represents the number of shoulders measured in each study.

For studies that reported Bland-Altman limits of agreement (LOA) but did not report SD, standard deviation was recalculated using the formula:

$$SD = \frac{Upper\ LOA - Lower\ LOA}{3.92}$$

Random-effects meta-analyses were performed using the restricted maximum likelihood method to account for between-study heterogeneity. Subgroup analyses were conducted based on motion type (flexion, abduction, extension, external

rotation [ER], and internal rotation [IR]), technology category (IMU vs. image-based), participant characteristics (healthy vs. pathological shoulders), and data acquisition methods (by patients vs. by assessors). Heterogeneity was assessed using the I^2 statistic. Differences between subgroups were assessed using a chi-squared (Q) test for subgroup differences, as implemented in the random-effects meta-analysis model.

Where substantial heterogeneity was identified ($I^2 > 50\%$), potential sources of heterogeneity were explored through subgroup analyses based on clinically relevant variables, including motion type, technology type, population characteristics, and data acquisition methods. Formal meta-regression was not performed due to the limited number of studies included in the quantitative synthesis. A 95% prediction interval were calculated to estimate the expected range of true effects in future studies, offering a more clinically meaningful interpretation of variability.

A funnel plot and Egger's regression test were used to assess publication bias. Statistical significance was defined as $p<0.05$. In accordance with PRISMA guidance, included studies were also assessed for potential meta-bias through visual inspection of funnel plot asymmetry and consideration of study size and reporting patterns. No formal statistical correction for publication bias was applied.

Reliability outcomes were assessed using intraclass correlation coefficients (ICCs), which quantify the consistency of measurements across repeated assessments or between raters. ICC values were interpreted in the context of measurement reliability, with higher values indicating greater agreement. Where multiple ICC values were reported within a study, these were summarised in Table 1 to maintain clarity and readability, with full values and confidence intervals considered during data extraction.

The overall confidence in the cumulative evidence was considered in relation to study quality, consistency of findings, and risk of bias. A formal GRADE assessment was not undertaken due to the methodological heterogeneity and diversity of outcome measures across included studies; however, key domains relevant to evidence certainty were considered in the interpretation of findings. [17]

Where quantitative synthesis was not appropriate due to heterogeneity in study design, outcome reporting, or measurement methods, a structured narrative synthesis was undertaken. This approach is consistent with Cochrane guidance, which recommends narrative or semi-quantitative synthesis when statistical pooling is not feasible or meaningful. [6, 20, 42]

The narrative synthesis aimed to provide a transparent and systematic summary of study findings, focusing on patterns in measurement approaches, reported accuracy and reliability metrics, and methodological differences across studies. This approach has been applied in previous reviews addressing heterogeneous intervention or measurement studies, allowing for an analytic rather than purely descriptive summary of the evidence.

Ethics

The methodology of this systematic review involves the review and synthesis of pre-existing published data, thereby making this phase of the study exempt from ethical approval.

Results

The results are presented in three stages: (1) an initial description of included studies and their characteristics, (2) a narrative synthesis of studies not included in the quantitative meta-analysis, and (3) quantitative synthesis of accuracy outcomes using meta-analysis, followed by assessment of reliability and risk of bias.

Study Selection and Characteristics

A total of 26 studies met the inclusion criteria and were analysed in this review. Of these, 17 studies focused on the accuracy of remote shoulder ROM measurement tools and 10 were able to be included in the meta-analysis. Studies examined reliability were narratively synthesised.

Of the 17 studies assessing accuracy, 7 were not included in the meta-analysis due to methodological or reporting limitations, including lack of appropriate agreement data (e.g. absence of mean bias or limits of agreement), reporting of compound or non-isolated movements, or focus on outcomes not directly aligned with ROM accuracy (e.g. strength or algorithm development without reference comparison).

The included studies utilised a range of remote measurement approaches, most commonly smartphone-based applications and wearable inertial measurement units (IMUs), with comparators including handheld goniometers, digital inclinometers, optical motion capture systems, and in-person clinician assessment.

The studies varied in methodologies, sample sizes, and comparator tools, covering a broad spectrum of remote measurement approaches, including smartphone applications, wearable inertial measurement units (IMUs), and markerless motion capture systems. The overall study characteristics are summarised in **Table 1.**

Table 1. Overall study characteristics summary. Abbreviations: ICC = intraclass correlation coefficient (measure of reliability for continuous data); κ = kappa coefficient (measure of agreement for categorical data); r = correlation coefficient (measure of linear association between variables); ROM = range of motion; IMU = inertial measurement unit. ICC values are reported as measures of test-retest or inter-rater reliability with individual 95% confidence interval, while κ values reflect agreement beyond chance for categorical assessments. KR-20 (Kuder–Richardson reliability) measures internal consistency reliability used for assessments with dichotomous items. Correlation coefficients (r-values) indicate the strength of association between measurement methods but do not represent agreement.

Author/ Year	Study Design	Population	Mean Age (years ± SD)	No. of participants	Number of Shoulders Assessed	Remote ROM Method	Comparators	Shoulder Motion Metrics Studied	Validity	Reliability	Test-Retest Reliability/ Intra-rater Reliability	Inter-rater Reliability	Reliability among compared instruments
Bechard et al. (2020) [2]	Measurement validity and reliability study	Healthy	Not reported	32	64	Force sensing device (FoSe)	Hand-Held dynamometer	Strength	Yes	Yes	N/A	N/A	ICC: ER: 0.74 (0.55–0.86) Left; 0.74 (0.56–0.87) Right Shoulder IR: 0.89 (0.80–0.95) Left; 0.73 (0.54–0.86) Right
Boissy et al. (2017) [3]	Measurement validity study	Healthy	30.9± 10.1	25	25	Digital Goniometer	Optic motion capturing system	Abduction, Flexion, ER	Yes	No			

Borresen et al. (2023) [4]	Measurement reliability study	Pathological	Not reported	15	15	Augmented Reality Telerehabilitation System with Haptics	In-person assessment	Elevation, Depression, IR, ER, Adduction, Abduction, Protraction, Retraction	No	Yes	N/A	κ = Elevation - 0.09 (-0.4 – 0.2) ER 0.25 (-0.2 – 0.7) Adduction - 0.11 (-0.3 – 0.0) Abduction 0.44 (-0.1 – 1.0) Protraction 0.44 (-0.1 – 1.0) Retraction 0.15 (-0.1 – 0.4)	N/A
Chan et al. (2022) [7]	Measurement validity study	Healthy	25 (SD not reported)	19	19	Wearable IMU	Optic motion capturing system	Flexion, Extension, ER, Abduction	Yes	No			

Chen et al. (2020) [8]	Measurement reliability study	Healthy & Pathological	53.0 ± 6.2 and 56.1 ± 13.3 (two groups)	25	25	IMU + Mobile App	Hand-Held Goniometer	Flexion, Extension, Abduction, IR, ER	No	Yes	N/A	ICC: Abduction: 0.97 (0.95–0.98) active; 0.98 (0.96–0.99) passive Flexion: 0.95 (0.92–0.97) active; 0.90 (0.83–0.94) passive Extension: 0.77 (0.64–0.87) active; 0.80 (0.68–0.89) passive ER: 0.95 (0.92–0.97) active; 0.96 (0.93–0.98) passive IR: 0.91 (0.86–0.95) active; 0.97 (0.94–0.98) passive	N/A

Cui et al. (2019) [12]	Measurement reliability study	Pathological	56.3 (SD not reported)	25	25	IMU + Virtual Reality	Hand-Held Goniometer	Flexion, Abduction, IR, ER	No	Yes	N/A	N/A	r-values: Flexion: 0.997 Abduction: 0.978 IR: 0.984 ER: 0.897
Cuesta-Vargas et al. (2016) [11]	Measurement validity and reliability study	Healthy & Pathological	Not reported	37	37	Mobile App	IMU	Abduction	Yes	Yes	ICC: Healthy: 0.78 (0.40 – 0.93) Pathological: 0.98 (0.94 – 0.99)	ICC: Healthy: 0.49 (0.08 – 0.82) Pathological: 0.99 (0.98 – 1.00)	N/A
Çubukçu et al. (2020) [10]	Measurement validity and reliability study	Healthy	22.1± 3.1	40	80	Kinect V2	Hand-Held & Digital Goniometer	Abduction, Flexion, Extension, IR, ER	Yes	Yes	ICC (no CI reported): Abduction 0.86 Flexion: 0.85 Extension: 0.62 IR: 0.97 ER: 0.87	N/A	N/A
Gushikem et al. (2022) [16]	Measurement validity and reliability study	Pathological	29.3 (SD not reported)	21	21	Tele-assessment	In-person assessment	Abduction, Strength	Yes	Yes	ICC: Abduction 0.87 (0.69 – 0.95)	N/A	N/A

Hwang et al. (2023) [21]	Measurement validity and reliability study	Healthy & Pathological	35.2 (SD not reported)	10 Healthy, 10 Pathological	20	Surgeon Estimation	Digital Goniometer	Flexion, Abduction, ER, IR	Yes	Yes	N/A	ICC Healthy: Flexion: 0.93 (0.86–0.98); ER 0.90 (0.81–0.97); IR 0.77 (0.60–0.92). Pathological: Flexion 0.88 (0.77–0.96); ER 0.87 (0.75–0.96); IR 0.74 (0.57–0.91).	N/A
Jayaraman et al. (2020) [22]	Measurement validity study	Healthy	28 ± 4 (Phase 1) / 30 ± 3 (Phase 2)	75	75	Apple Watch	Hand-Held Goniometer	Abduction, Flexion, Extension	Yes	No			
Lam et al. (2024) [27]	Measurement validity and reliability study	Healthy	28.9± 11.8	30	30	Markerless Motion Capture System with iPad Pro	Hand-Held Goniometer	Abduction, Flexion	Yes	Yes	ICC (95% CI not reported) (Static ranges): 0.17 to 0.80; ICC (Tasks ranges): 0.57 to 0.96	N/A	N/A
Niu et al. (2024) [30]	Algorithmic optimisation & tracking accuracy study	Healthy	Not reported	16	16	Wearable smartwatch	Vicon Camera	Compound movements	No	No			

Ongvisatepaiboon et al. (2016) [31]	Quasi-experimental study	Healthy	Not reported	20	20	Audio-biofeedback (ABF) system	Non-ABF system	Flexion	No	No			
Pereira et al. (2023) [33]	Measurement validity study	Healthy	Not reported	15	15	Mobile App using pose estimation	Qualisys Motion Capture System (QTM)	Compound movements	Yes	No			
Rajkumar et al. (2021) [34]	Measurement reliability study	Healthy	Not reported	17	34	Wearable IMU for exergames (WISE)	Kinect	Flexion-Extension, IR-ER, Abduction-Adduction (Compounded)	Yes	Yes	ICC (95% CI not reported): 0.46 to 0.90	N/A	ICC (95% CI not reported): 0.57 to 0.81
Ramkumar et al. (2018) [35]	Measurement validity study	Healthy	27.0 (SD not reported)	10	10	Smartphone App	Hand-Held Goniometer	Flexion, Abduction, IR, ER	Yes	No			
Roldán-Jiménez et al. (2019) [38]	Measurement reliability study	Healthy & Pathological	Not reported	16 Healthy, 27 Pathological	43	Smartphone in-built IMU	Wearable IMU sensors	Abduction, Flexion	No	Yes	N/A	N/A	ICC (Abduction): 0.86 to 0.97 (95% CI 0.73–0.99); ICC (flexoextension axis): 0.48 to 0.87 (95% CI 0.41–0.93); ICC (rotation axis): 0.43 to 0.62 (95% CI 0.10–0.80).

Sahu et al. (2022) [41]	Measurement validity and reliability study	Healthy & Pathological	Not reported	24 Healthy, 16 Pathological	40	On-Screen App	Hand-Held Goniometer	Abduction, Flexion, IR, ER	Yes	Yes	ICC (consultant measurements): Abduction 0.97 (0.86- 0.99); Flexion 0.99 (0.97 – 0.99); IR 0.97 (0.94 – 0.99); ER 0.98 (0.94 – 1.0)	ICC: Abduction 0.86 (0.67 – 0.94); Flexion 0.94 (0.85 – 0.97); IR 0.95 (0.88 – 0.98); ER 0.97 (0.94 – 0.99).	
Seo et al. (2016) [43]	Measurement validity study	Healthy	28.5 (SD not reported)	10	10	Different Kinect sensor placements	3D Motion Capture	Flexion-Extension, IR-ER, Abduction-Adduction (Compounded)	Yes	No			
Shimizu et al. (2022) [46]	Measurement validity and reliability study	Healthy	Not reported	19	19	Smartphone App	Hand-Held Goniometer	Abduction, Flexion, Extension, IR, ER	Yes	Yes	ICC: Abduction 0.95 (0.91 – 0.98); Flexion 0.94 (0.88 – 0.98); Extension 0.79 (0.62 – 0.91); IR 0.91 (0.82 – 0.96); ER 0.93 (0.86 – 0.97).	N/A	N/A

Soeters et al. (2023) [47]	Measurement validity and reliability study	Healthy	31.4± 11.7	30	60	Smartphone App	Hand-Held Goniometer	Abduction, Flexion, Extension, ER, IR	Yes	Yes	ICC: Abduction 0.95 (0.93 – 0.97); Flexion 0.94 (0.90 – 0.96); Extension 0.90 (0.84 – 0.93); ER 0.93 (0.88 – 0.95); IR 0.75 (0.62 – 0.84).	ICC: Abduction 0.90 (0.85 – 0.94); Flexion 0.96 (0.94 – 0.98); Extension 0.96 (0.93 – 0.97); ER 0.95 (0.92-0.97); IR 0.93 (0.89 - 0.95).	N/A
Tozawa et al. (2023) [50]	Measurement reliability study	Healthy	Not reported	16	16	Zoom & Smartphone App	N/A	Flexion	No	Yes	ICC: Examiner 1: 0.91 (0.81 – 0.97) Examiner 2: 0.97 (0.93 – 0.99)	ICC: 0.95 (0.85 – 0.98)	N/A
Wang et al. (2020) [51]	Measurement validity study	Healthy & Pathological	Not reported	30 Healthy, 1 Pathological	31	Kinect	N/A	Flexion, Abduction, Rotation	Yes	No			

Wang et al. (2022) [52]	Measurement reliability study	Pathological	50.2± 16.2	32	32	Telemedicine examination	In-person assessment	Abduction, Adduction, Flexion, IR, ER, Strength	No	Yes	N/A	KR-20: Abduction 0.78 Adduction 0.88 Flexion 0.37 IR: 0.66 ER: 0.52 Strength: 0.57	KR-20: Abduction: 0.50 Adduction: 0.32 Flexion: 0.51 IR: 0.29 ER: 0.44 Strength: 0.37
Wang et al. (2023) [53]	Measurement reliability study	Healthy	Not reported	25	25	Webcam based machine learning approach (Blazepose)	Optic motion capturing system (OptiTrack)	Flexion-Extension (Sagittal plane), Abduction-Adduction (Frontal plane)	No	Yes	ICC: Flexion 0.94 (0.90 – 0.96); Extension 0.93 (0.89 – 0.96); Abduction 0.97 (0.95 – 0.98); Adduction 0.92 (0.87 – 0.95).	N/A	ICC: Flexion 0.18 (-0.14 – 0.40); Extension 0.92 (0.88 – 0.94); Abduction 0.68 (0.56 – 0.77); Adduction 0.89 (0.84 – 0.92)

Overall study quality was mixed. Assessment of reporting bias revealed asymmetry in the funnel plot and a statistically significant result from Egger's test ($p = 0.0051$), suggesting the presence of publication bias. These methodological limitations should be considered when interpreting the pooled meta-analytic results.

A notable characteristic of the included studies is the predominance of young, healthy volunteer populations. Most studies excluded individuals with musculoskeletal pathology, comorbidities, or functional impairments. As a result, the generalisability of the findings to older adults or those with shoulder conditions is limited.

Populations included both healthy individuals and patients with pathological shoulder conditions such as rotator cuff tears, adhesive capsulitis, and post-operative rehabilitation cases. ROM assessments covered flexion, abduction, extension, internal rotation (IR), and external rotation (ER). The reference standards commonly used were goniometry, optical motion capture, and clinician assessments.

Studies evaluating accuracy but not included in the meta-analysis

Seven studies evaluating the accuracy of remote shoulder assessment methods were not included from the meta-analysis but narratively synthesised due to methodological or reporting limitations but were nonetheless important for understanding the broader context of remote measurement approaches.

Two of these studies, by Bechard and colleagues (2020) and Gushikem and colleagues (2022), focused solely on strength assessments rather than range of motion (ROM), which precluded their inclusion in a meta-analysis specifically targeting ROM measurement agreement. They do, however, provide broad insights in understanding important quality of shoulder motion metrics in addition to range of motion alone. [2, 16] Similarly, Niu and colleagues (2024) did not report any quantitative data on accuracy or reliability. Instead, it emphasised technical advancements in real-time motion tracking using ultrasound and inertial sensors, prioritising system development and algorithmic accuracy without comparing against established reference standards or applying statistical tools like Bland-Altman analysis. Although not contributing directly to the pooled analysis, this study offers valuable insights into emerging technologies. [30]

Cuesta-Vargas and colleagues provided ROM data using both inertial sensors and smartphone-based methods in healthy and pathological populations. In healthy participants, mean ± SD values were 169.07° ± 4.96° using smartphone measurements and 154.22° ± 19.27° using inertial sensors. In pathological participants, corresponding values were 93.54° ± 40.88° and 87.86° ± 47.41°, respectively. The study, however, lacked the required paired differences between measurements for appropriate Bland-Altman analysis, instead reporting only mean and standard deviation values for separate groups. Without direct within-subject comparisons, the assumptions required to estimate limits of agreement may not be valid, making the results unsuitable for meta-analysis. [11]

Similarly, Rajkumar and colleagues (2021) reported data from a Bland-Altman analysis comparing Kinect and WISE systems but presented flexion-extension as a combined metric. Mean root mean square error (RMSE) values ranged from 7.72° ± 2.82° to 11.29° ± 3.10° across exercises, with lower errors observed following

dynamic time warping (6.28° ± 2.57° to 9.62° ± 3.88°). Because the meta-analysis required individual movement assessments (e.g., flexion and extension separately), this aggregation limited compatibility with pooled estimates. [34]

Pereira and colleagues (2023) assessed compound shoulder movements rather than isolated ROM parameters, comparing mobile app-based pose estimation, motion capture (QTM), and DTW algorithms. the mobile app-based pose estimation demonstrated a mean score of 94.08 ± 1.18, indicating high consistency in movement pattern detection. In contrast, DTW method showed greater variability, with a mean of 53.74 ± 25.59. The reference QTM system demonstrated wider dispersion (mean 78.8 ± 21.61), reflecting sensitivity to variations in movement execution. The study found wide variability in agreement between methods, especially between DTW and QTM, but its focus on multidimensional movements prevented integration with studies measuring isolated joint angles. [33]

Lastly, Seo and colleagues (2016) offered critical contextual information, showing that Kinect may systematically overestimate ROM due to depth perception limitations. For shoulder elevation, the mean ROM error was 5° ± 1°, representing the lowest error among the assessed joints. Across sensor configurations, the lowest overall error was observed when the Kinect was positioned at 45° elevation in front of the subject (10° ± 1°), whereas the conventional frontal placement resulted in higher error (22° ± 2°), and the greatest error was observed at contralateral positioning (38° ± 2°).While this does not lend itself to quantitative synthesis, it offers a plausible explanation for overestimation trends observed in other remote ROM studies and is therefore valuable in interpreting pooled results. [43]

Accuracy of Remote Shoulder ROM Measurement

The meta-analysis included studies that quantitatively assessed the accuracy of remote ROM measurement technologies against reference standards. Four separate subgroup meta-analyses were performed:

1. By motion type
2. By technology type
3. By population (healthy vs. pathological)
4. By data acquisition methods

Overall Accuracy Findings

Across all studies, the difference of the means was 2.63° (95% CI: 1.52, 3.74), indicating a small but consistent overestimation of shoulder ROM when measured remotely. There was significant heterogeneity (I^2 = 98.62%), reflecting variations in study methodologies and sample populations. The overestimation tendency was observed across all movement planes, with statistically significant overestimation primarily driven by external rotation (ER), flexion, and internal rotation (IR).

Accuracy by Motion Type (Figure 2)

When analysed by shoulder movement type, the difference of the means were as follows:

- **Abduction:** 2.54° (95% CI: -0.54, 5.63; I^2 = 98.24%), which was not significantly different from reference standard methods. The 95% prediction interval is − 9.22° to 14.30°.
- **External Rotation (ER):** 1.33° (95% CI: 0.10, 2.57; I^2 = 91.28%), showing a small but statistically significant difference. The 95% prediction interval is – 2.83° to 5.49°.
- **Extension:** -0.13° (95% CI: -0.52, 0.26; I^2 = 41.20%), which showed no significant difference between remote and reference standard methods. The 95% prediction interval is – 1.18° to 0.92°.
- **Flexion:** 3.82° (95% CI: 1.52, 6.12; I^2 = 98.06%), indicating significant overestimation. The 95% prediction interval is – 5.37° to 13.01°.

- **Internal Rotation (IR):** 4.14° (95% CI: 1.71, 6.57; I^2 = 94.94%), showing the largest measurement discrepancy. The 95% prediction interval is – 4.31° to 12.59°.

Notably, remote measurements consistently overestimated ROM across most movements, except for abduction and extension, where no systematic over- or underestimation was observed.

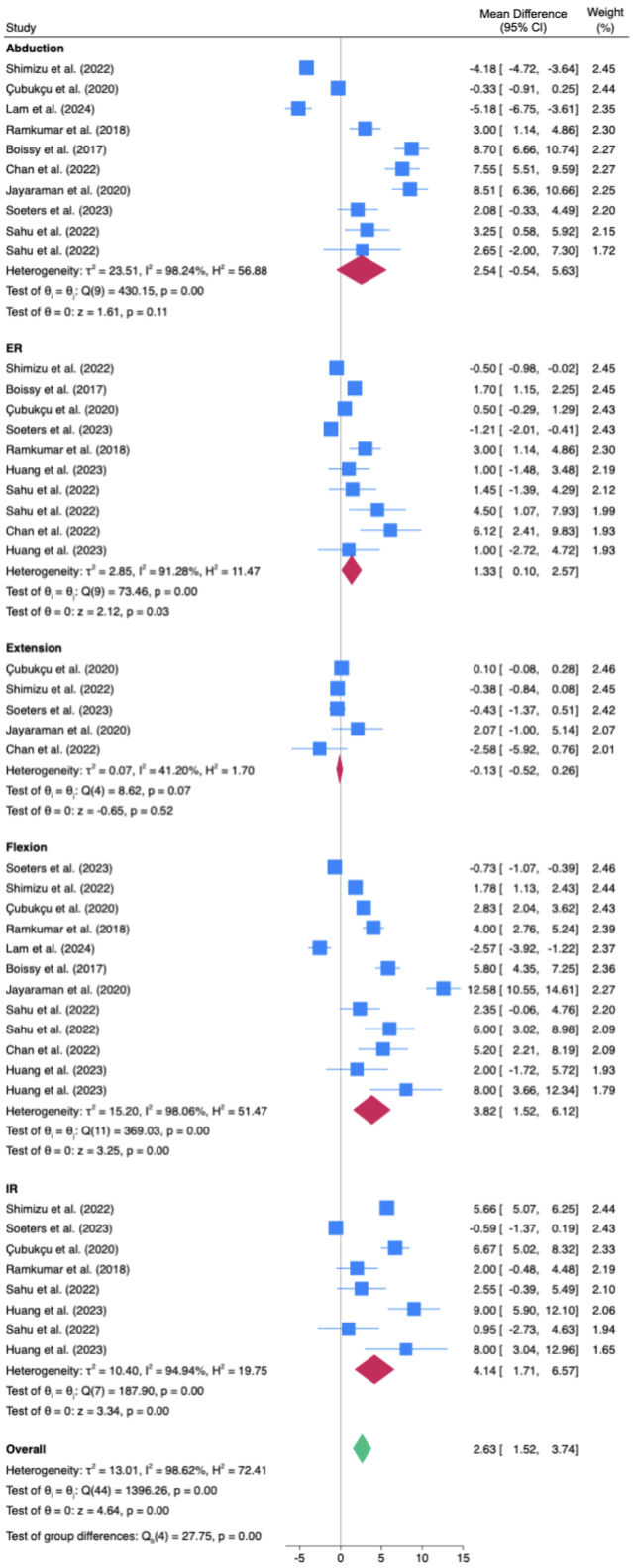

Figure 2. Forest plot of the meta-analysis of accuracy of remote shoulder assessment by motion type. Mean differences (in degrees) between remote and reference measurements are presented as point estimates (dots), with 95% confidence intervals (horizontal bars). Negative values indicate underestimation by the remote method, while positive values indicate overestimation. Statistical heterogeneity is reported using the I^2 statistic.

Accuracy by technology types (Figure 3)

- **IMU subgroup**: The difference of the means was 2.10° (95% CI: 0.64°, 3.56°; I^2 = 98.55%), suggesting that, on average, IMU-based methods slightly overestimate shoulder ROM when compared to reference standards. This result was statistically significant. The 95% prediction interval is – 5.08° to 10.05°.
- **Non-IMU (Image-based technologies) subgroup**: The difference of the means was 3.13° (95% CI: 1.47°, 4.79°; I^2 = 97.65%), also indicating an overestimation with non-IMU methods, again statistically significant. The 95% prediction interval is – 6.05° to 12.31°.

The test of subgroup differences yielded a p-value of 0.36, indicating that there was no statistically significant difference between IMU and non-IMU methods in terms of difference of the means.

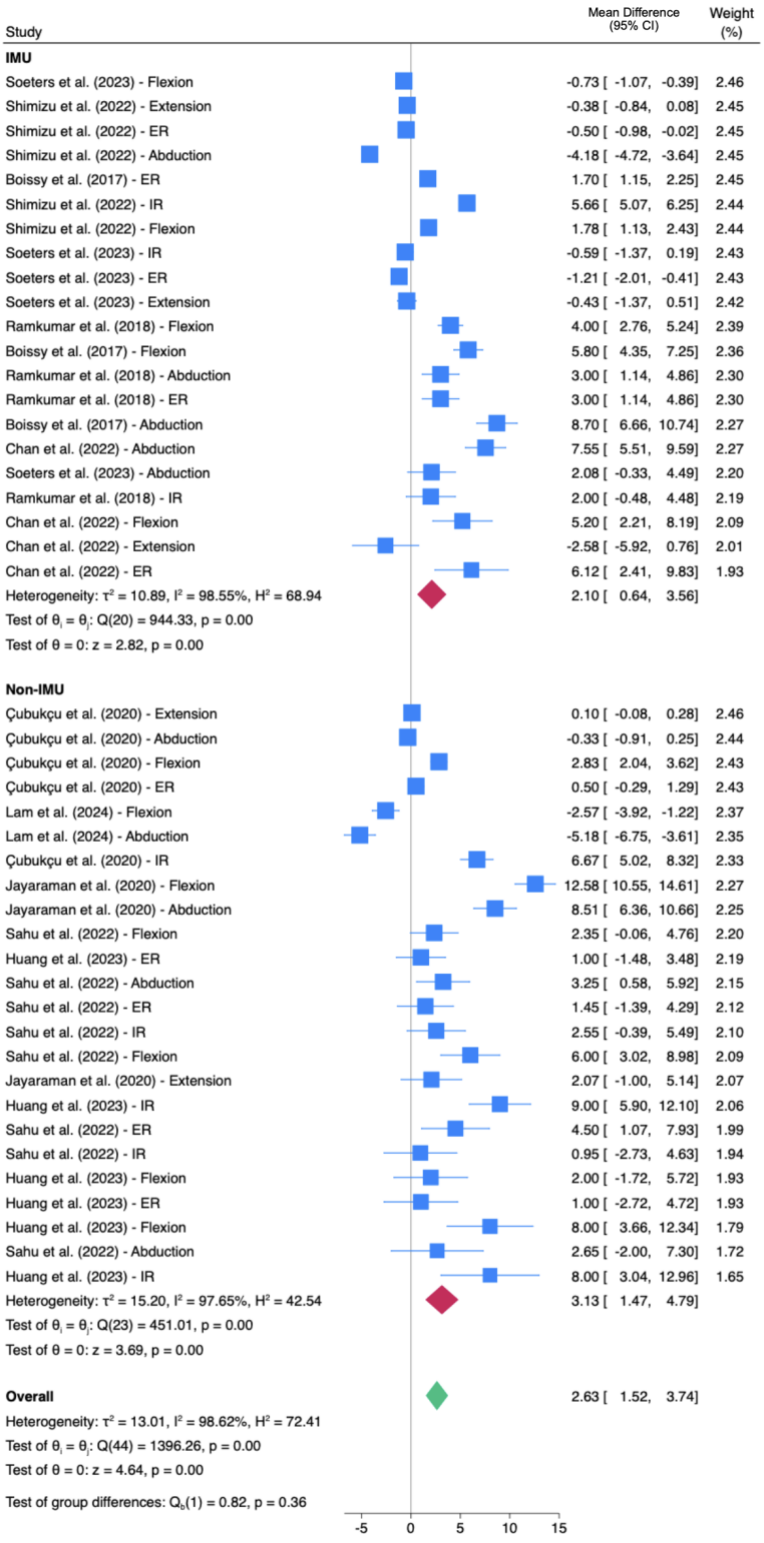

Figure 3. Forest plot of the meta-analysis of accuracy of remote shoulder assessment by technology types (IMU versus non-IMU). Mean differences (in degrees) between remote and reference measurements are presented as point estimates (dots), with 95% confidence intervals (horizontal bars). Negative values indicate underestimation by the remote method, while positive values indicate overestimation. Statistical heterogeneity is reported using the I^2 statistic.

Accuracy in Healthy vs. Pathological Shoulders (Figure 4)

- Healthy participants had a difference of the means of 2.37° (95% CI: 1.15, 3.59; I^2 = 98.89%), suggesting overestimation. The 95% prediction interval is – 5.54° to 10.28°.
- Pathological shoulders exhibited a difference of the means of 4.33° (95% CI: 2.17, 6.48; I^2 = 53.72%), reflecting higher variability in measurement error. The 95% prediction interval is – 0.53° to 9.19°.

While the test of subgroup differences was non-significant (p = 0.12), suggesting no statistically significant difference in measurement between healthy and pathological populations, the observed differences was higher in the pathological group (4.33° vs. 2.37°).

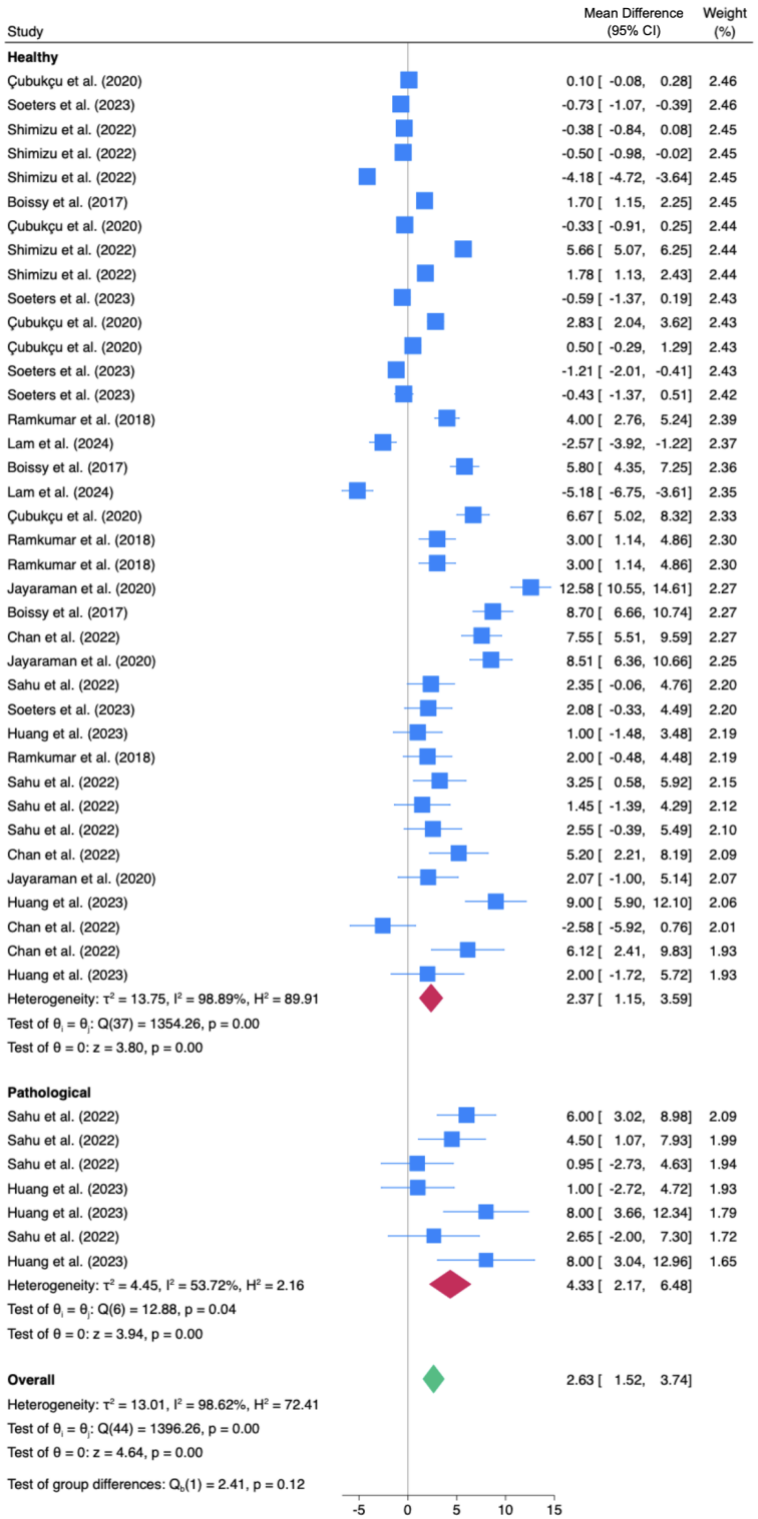

Figure 4. Forest plot of the meta-analysis of accuracy of remote shoulder assessment by population types (healthy vs. pathological). Mean differences (in degrees) between remote and reference measurements are presented as point estimates (dots), with 95% confidence intervals (horizontal bars). Negative values indicate underestimation by the remote method, while positive values indicate overestimation. Statistical heterogeneity is reported using the I^2 statistic.

Accuracy by data acquisition methods (Figure 5)

- For assessor-guided measurements, the difference of the means was 3.42° (95% CI: 2.16 to 4.68; I^2 = 97.33%), suggesting a statistically significant overestimation. The 95% prediction interval is – 4.80° to 11.64°.
- For self-measurements, the difference of the means was 0.12° (95% CI: -1.50 to 1.73; I^2 = 98.65%), which was not statistically significant. The 95% prediction interval is – 4.98° to 5.22°.

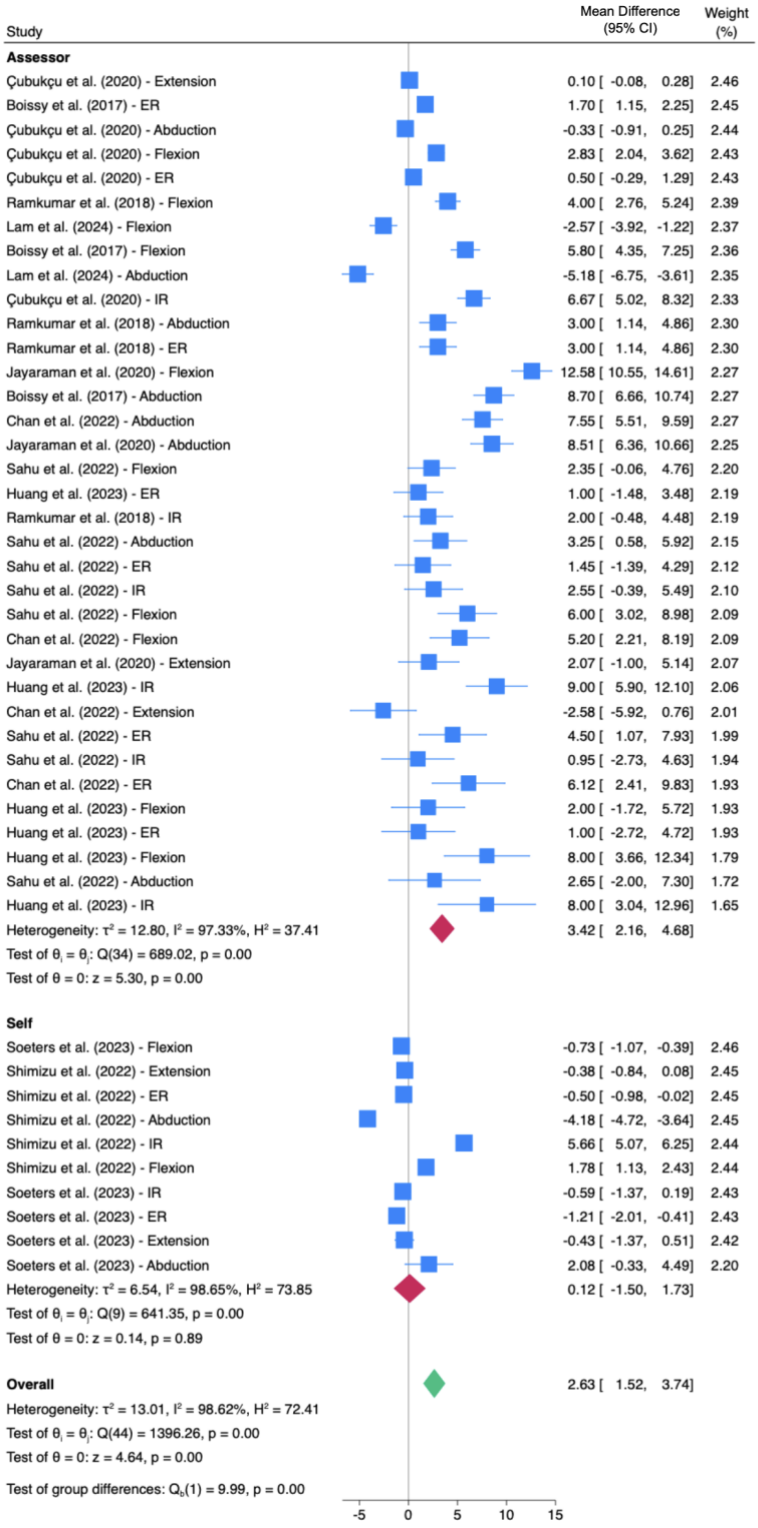

Figure 5. Forest plot of the meta-analysis of accuracy of remote shoulder assessment by data acquisition methods (assessor vs. self). Mean differences (in degrees) between remote and reference measurements are presented as point estimates (dots), with 95% confidence intervals (horizontal bars). Negative values indicate underestimation by the remote method, while positive values indicate overestimation. Statistical heterogeneity is reported using the I^2 statistic.

Publication Bias

The funnel plot (**Figure 6**) showed an asymmetric distribution of 45 datapoints reported from 10 studies, suggesting potential small-study effects. Egger's test confirmed a statistically significant bias (p = 0.0051), indicating the possibility of publication bias.

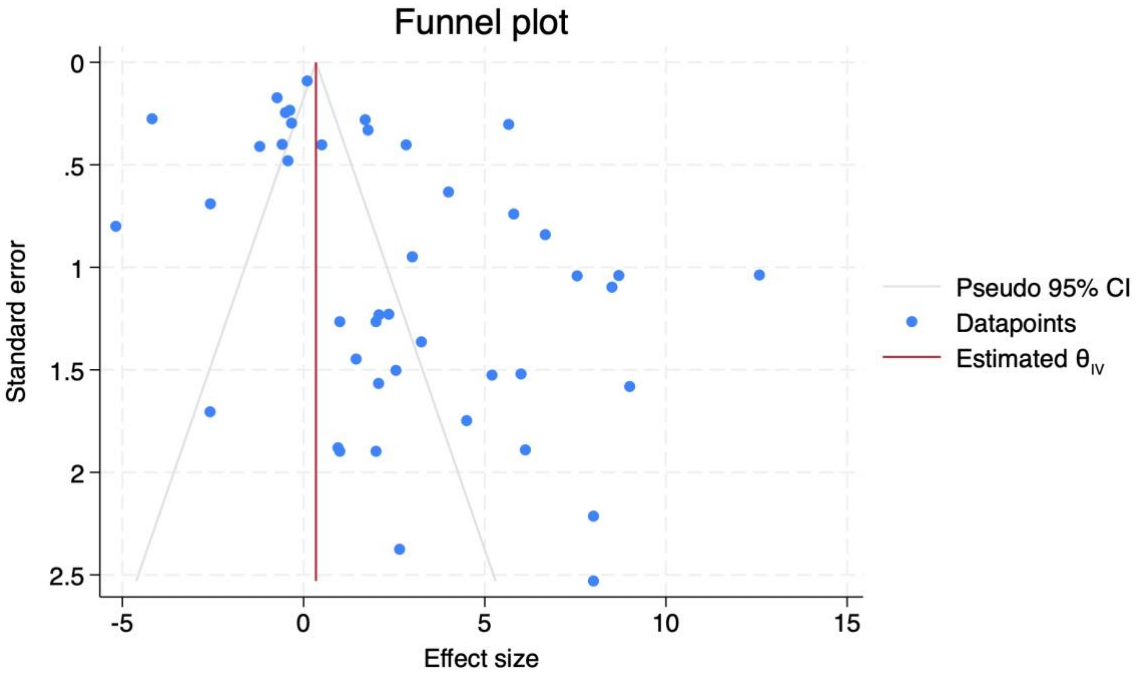

Figure 6. Funnel plot of the included data

Reliability of Remote Shoulder ROM Measurement

A total of 17 studies reported reliability data relevant to remote shoulder ROM measurement. Reliability was assessed using various methods, including test-retest, inter-rater, and inter-instrument comparisons, across both healthy and pathological populations.

Test-Retest Reliability

Reliability outcomes were reported using both intra-rater and test–retest reliability metrics. Intra-rater reliability refers to the consistency of measurements taken by the same assessor across repeated trials, whereas test–retest reliability reflects the stability of a measurement over time under similar conditions. Although these terms are sometimes used interchangeably in the literature, they represent distinct constructs and have been reported accordingly in this review.

Wang and colleagues (2023) reported test–retest reliability (referred to as intra-rater reliability in the study) using repeated measurements under the same conditions with a BlazePose-based machine learning system. ICCs range from 0.92 (95% CI 0.87 to 0.95) for shoulder adduction to 0.97 (95% CI 0.95 to 0.98) for shoulder abduction. [53] Sahu and colleagues (2022) also demonstrated high test-retest reliability in patients using a protractor app measured by consultant raters, with ICCs ranging from 0.97 (95% CI 0.86-0.99) for shoulder abduction to 0.99 (95% CI 0.97-0.99) for shoulder flexion. [40] Similarly, Çubukçu and colleagues (2020) found high intra-rater reliability with ICCs between 0.62 for shoulder extension and 0.97 for internal rotation using a Kinect V2 system, while Shimizu and colleagues (2022)observed ICCs varied depending on the range of movement from 0.79 (95% CI 0.62 to 0.91) for extension to 0.95 (95% CI 0.91-0.98) for abduction with a smartphone self-measurement tool. [10, 46]

Lam and colleagues (2024) however, highlighted greater variability in intra-rater reliability using a dual-iPad LiDAR setup, with ICCs ranging from 0.17 to 0.96 across static and functional tasks (95% CI not reported). [27] Tozawa and colleagues (2023) showed excellent intra-rater reliability for shoulder flexion between two separate examiners (ICC = 0.91–0.97). [50] Gushikem and colleagues (2022) reported intra-

rater ICCs of 0.87 (95% CI 0.69 – 0.95) for abduction strength assessments in patients with brachial plexus injury. [16]

Inter-Rater Reliability

Inter-rater reliability was reported across a range of settings. Soeters and colleagues (2023) reported excellent ICCs between 0.90 (95% CI 0.85 – 0.94) for abduction and 0.96 (95% CI 0.94 – 0.98) for flexion when comparing smartphone app and goniometer measures. [48] Tozawa and colleagues (2023) showed inter-rater ICC of 0.95 (95% CI 0.85 – 0.98) for shoulder flexion. [50] Chen and colleagues (2020) found inter-rater ICCs from 0.77 (95% CI 0.64 – 0.87) for active shoulder extension to 0.98 (95% CI 0.96 – 0.99) for passive shoulder abduction. [8]

In contrast, Borresen and colleagues (2023) reported no agreement to moderate agreement reporting with kappa statistics, with $\kappa = -0.11$ (95% CI -0.3 – 0.0) for shoulder adduction to 0.44 (95% CI -0.1 – 1.0) for shoulder abduction. Cuesta-Vargas and colleagues reported varied ICCs for inter-rater reliability ranging from 0.49 (95% CI 0.08 – 0.82) for healthy volunteers to 0.99 (95% CI 0.98 – 1.00) for shoulders with pathology present.

Agreement Between Instruments

Studies comparing remote measurement tools against reference standards or between different remote systems showed a range of agreement. Rajkumar and colleagues (2021) found ICCs between 0.57 and 0.81 (95% CI not reported) for various shoulder movements when comparing WISE and Kinect systems. [34] Roldán-Jiménez and colleagues (2019) reported excellent agreement between smartphone-based and dedicated IMU sensors for abductionICC 0.86 (95% CI 0.73-0.99), with lower values for internal rotation ICC 0.43 (95% CI 0.10-0.80). [38] Cui and colleagues (2019) demonstrated strong correlations from r = 0.897 for shoulder external rotation to r = 0.997 for shoulder flexion between wearable sensors integrated with virtual reality and therapist assessments. [12] Bechard and colleagues (2020) evaluated the agreement between a novel strength device and handheld dynamometry, reporting ICCs from 0.73 (95% CI 0.54-0.86) for right sided shoulder internal rotation to 0.89 (95% CI 0.80-0.95). [2] Wang and colleagues (2022) used KR-20 to evaluate dichotomous findings across telemedicine and in-person sessions, with varied

735 agreement values ranging from 0.29 for internal rotation to 0.51 (no 95% CI
736 reported) for flexion depending on the motion type and method. [52]

Risk of Bias Assessment

Risk of bias was assessed using the QUADAS-2 tool, with pre-defined signalling questions tailored to the context of accuracy and reliability of instruments in remote shoulder range of motion (ROM) measurements (See **Appendix 2**). A visual summary of the QUADAS-2 domain assessments is shown in **Figure 7**, and detailed study-level ratings are available in **Table 2**.

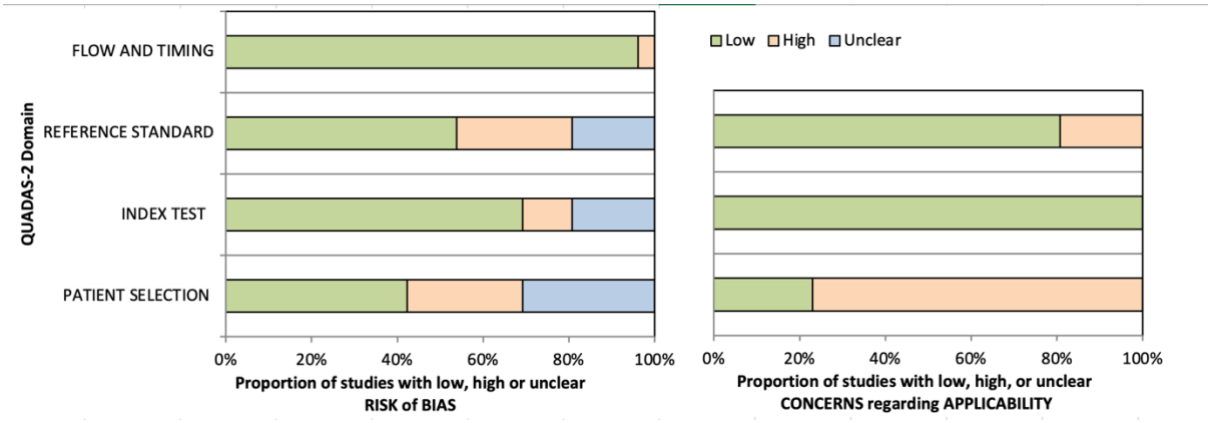

Figure 7. Visual Representation of Risk of Bias using QUADAS-2 Reporting Tool

747 **Table 2. Tabular presentation for QUADAS-2 results**

Study	RISK OF BIAS				APPLICABILITY CONCERNS		
	PATIENT SELECTION	INDEX TEST	REFERENCE STANDARD	FLOW AND TIMING	PATIENT SELECTION	INDEX TEST	REFERENCE STANDARD
Bechard et al. (2020)	?	☺	?	☺	☹	☺	☺
Boissy et al. (2017)	?	☹	?	☺	☹	☺	☺
Borresen et al. (2023)	☺	☺	☺	☺	☺	☺	☺
Chan et al. (2022)	☹	☹	☹	☺	☹	☺	☺
Chen et al. (2020)	☹	☹	☺	☺	☹	☺	☺
Cui et al. (2019)	☹	☺	☺	☺	☹	☺	☺
Cuesta-Vargas et al. (2016)	☺	?	☺	☺	☺	☺	☺
Çubukçu et al. (2020)	☺	☺	☹	☺	☹	☺	☺
Gushikem et al. (2022)	?	☺	☹	☺	☺	☺	☹
Hwang et al. (2023)	☹	☺	☺	☺	☹	☺	☺

Jayaraman et al. (2020)	?	?	☹	☺	☹	☺	☺
Lam et al. (2024)	☺	?	?	☺	☹	☺	☺
Niu et al. (2024)	?	?	?	☺	☹	☺	☺
Ongvisatepaiboon et al. (2016)	?	☺	☹	?	☹	☺	☹
Pereira et al. (2023)	☹	☺	☹	☺	☹	☺	☹
Rajkumar et al. (2021)	?	☺	☹	☺	☹	☺	☹
Ramkumar et al. (2018)	☺	☺	☺	☺	☹	☺	☺
Roldán-Jiménez et al. (2019)	☺	☺	☺	☺	☺	☺	☹
Sahu et al. (2022)	☹	☺	☺	☺	☺	☺	☺
Seo et al. (2016)	☺	☺	☺	☺	☹	☺	☺
Shimizu et al. (2022)	☹	☺	☺	☺	☹	☺	☺
Soeters et al. (2023)	☺	☺	☺	☺	☹	☺	☺
Tozawa et al. (2023)	☺	☺	☺	☺	☹	☺	☺
Wang et al. (2020)	?	?	?	☺	☹	☺	☺

Wang et al. (2022)	☺	☺	☺	☺	☺	☺	☺
Wang et al. (2023)	☺	☺	☺	☺	☹	☺	☺

748 ☺Low Risk ☹High Risk ? Unclear Risk

Out of 26 studies included, patient selection was the most frequent source of bias, with the majority of studies (approximately 60%) rated as high or unclear risk in this domain. This was primarily due to the predominance of healthy volunteer cohorts, limiting applicability to clinical populations with shoulder pathology.

The index test domain demonstrated generally low risk of bias, although approximately 25% of studies had unclear methods of test administration or lacked sufficient detail on blinding procedures. The reference standard domain also showed moderate concerns, with some studies failing to provide adequate justification or validation for the gold-standard comparator. Flow and timing were mostly well handled, with nearly all studies administering both index and reference tests within a clinically appropriate timeframe.

Applicability concerns were similarly concentrated in patient selection, where around 80% of studies showed high concern due to unrepresentative populations. In contrast, applicability concerns for the index test and reference standard were minimal across the dataset.

Discussions

This systematic review and meta-analysis evaluated the accuracy and reliability of remote shoulder ROM measurement methods. The results suggest that existing remote measurements consistently overestimated ROM. While many remote measurement technologies demonstrated acceptable reliability, their agreement with reference standards was variable.

The overall mean bias of 2.63° (95% CI: 1.52, 3.74) indicates that remote ROM measurements tend to slightly overestimate shoulder motion compared to reference-standard assessments **(Figure 2)**. This overestimation was most notable in external rotation, flexion, and internal rotation, while extension showed the smallest discrepancy. Due to the substantial heterogeneity across studies (I^2 = 98.62%, τ^2 = 13.01), we also calculated a 95% prediction interval to better reflect the expected range of effects in future applications. The resulting interval (−5.79° to 11.05°) reveals considerable variability, including the possibility of underestimation in some contexts. This wide range highlights the influence of differing populations, methodologies, and technologies, and calls for caution in generalising the pooled estimate. An alternative interpretation of these findings is that the variability observed across studies may reflect limitations in the accuracy and precision of the technologies themselves, rather than solely differences in study design or methodology. The presence of wide limits of agreement and large prediction intervals suggests that, in some cases, remote measurement systems may produce inconsistent or imprecise estimates of shoulder range of motion.

Although the mean bias was statistically significant, both the pooled estimate and the prediction interval fall below the minimum clinically important difference (MCID) for glenohumeral joint ROM, which ranges from 11° to 16° when assessed by a single evaluator, and 14° to 24° when assessed by two. This suggests that, while measurable, the observed bias may not be clinically meaningful in most cases. [29]

Both IMU and non-IMU image-based motion tracking systems demonstrate a statistically significant overestimation of ROM, the magnitude of bias remains relatively small **(Figure 3)**: Sensor drift and recalibration issues – Over prolonged

use, IMU sensors may accumulate minor errors, leading to inconsistent readings; The findings of Seo and colleagues (2016) demonstrate that Kinect-based shoulder ROM measurement can achieve relatively low error under optimal conditions (~5°), although accuracy is highly dependent on sensor positioning. [43] This variability is consistent with the wide prediction intervals observed in the present meta-analysis and highlights the importance of methodological factors in influencing measurement error. Placement inconsistencies – Small variations in sensor positioning between measurement sessions or between different raters can introduce additional error; Soft tissue movement artifacts – Unlike rigidly mounted motion capture markers, IMU sensors placed on the skin may shift slightly due to soft tissue deformation and muscle movement, leading to variations in recorded ROM. [14, 19, 39] Importantly, these biases fall below the MCID thresholds. Therefore, while measurement variability exists, remote ROM assessments using either method are unlikely to introduce clinically meaningful discrepancies in routine monitoring.

Although pathological shoulders demonstrated a larger mean measurement difference compared to healthy participants, the test of subgroup differences was not statistically significant **(Figure 4)**. Therefore, any apparent increase in variability should be interpreted with caution and may reflect differences in data distribution rather than a true subgroup effect.

One possible explanation is that pathological shoulder conditions introduce greater variability in movement execution, due to factors such as pain, restricted range, compensatory movement patterns, and reduced motor control. These factors may lead to less consistent positioning and movement during assessment, increasing measurement variability. Alternatively, the observed variability may reflect limitations in the accuracy and precision of the measurement technologies themselves when applied to pathological populations. For example, altered movement patterns or reduced joint range may challenge tracking algorithms or sensor alignment, potentially reducing measurement accuracy. Findings from Cuesta-Vargas and colleagues (2016) support the variability observed in the present meta-analysis. Substantial differences in mean arm abduction angle and variability were observed between smartphone and inertial sensor measurements, particularly in pathological populations. [11]

Interestingly, subgroup analysis based on who performed the ROM measurement revealed that assessor-guided measurements (n = 8 studies) had a pooled mean bias of 3.42° (95% CI: 2.16 to 4.68), indicating statistically significant overestimation **(Figure 5)**. In contrast, self-measurements (n = 2 studies) demonstrated a much smaller and non-significant bias of 0.12° (95% CI: -1.50 to 1.73). While these findings suggest that self-measurement—when properly guided—may yield comparable accuracy with tighter agreement, the limited number of self-assessment studies precludes definitive conclusions. Nonetheless, this highlights the potential of patient-led measurements in remote settings, warranting further investigation in future research.

17 studies assessed the reliability of remote shoulder ROM measurements, encompassing test-retest, inter-rater, and inter-instrument comparisons across healthy and pathological populations. The wide variation in inter-rater ICC values across studies has important clinical implications. High ICC values (e.g. >0.80) indicate good agreement between assessors and suggest that remote measurement tools may be reliable in controlled or optimised conditions. However, lower ICC values (e.g. <0.60) indicate reduced agreement and raise concerns regarding consistency between users. In a clinical context, this variability suggests that measurements obtained using remote technologies may differ depending on the assessor, which could impact clinical decision-making, monitoring of progress, and treatment planning. This inconsistency may limit the reliability of these technologies in routine practice, particularly in settings where multiple clinicians or patients are involved in measurement, such as tele-rehabilitation or self-assessment contexts.

The observed variability in ICC values may be attributed to several factors, including differences in assessor experience, variability in patient movement execution, and sensitivity of the technology to positioning or environmental conditions. These factors may disproportionately affect remote assessment tools compared to traditional in-person measurement methods.

It is important to distinguish between intraclass correlation coefficients (ICC) and kappa statistics, as they assess different types of agreement. ICC is used for

continuous data and evaluates the consistency of quantitative measurements between raters or across time. In contrast, kappa statistics assess agreement for categorical data, such as classification of movement quality or presence/absence of impairment. As these metrics evaluate different constructs, differences in reported values between ICC and kappa are expected and do not represent conflicting results. Consequently, studies reporting kappa values may reflect agreement in clinical classification, whereas ICC values provide insight into measurement precision. Both are important but address different aspects of reliability.

From a clinical perspective, the variability in reliability metrics suggests that while remote technologies show promise, their consistency is not yet sufficient to fully replace standard measurement methods in all contexts. Clinicians should therefore interpret remote ROM measurements with caution, particularly when monitoring small changes over time or when multiple assessors are involved.

The QUADAS-2 assessment highlights several important limitations in the current evidence base. First, the predominance of healthy participants introduces a key source of bias and limits the applicability of findings to clinical populations. This interpretation is further supported by the subgroup analysis, which demonstrated a greater mean measurement difference in pathological shoulders (4.33°) compared to healthy participants (2.37°), although this difference did not reach statistical significance. While this finding should be interpreted with caution, it suggests a trend towards increased measurement discrepancy in pathological populations. Although remote measurement technologies may perform well under ideal conditions, their accuracy and reliability in patients with restricted or compensatory movement patterns remain uncertain. Second, inconsistencies in reporting or justifying the reference standards used—such as goniometers, optical motion capture, or clinician ratings—may have impacted the observed variability in agreement. In some studies, the reference standard itself was subject to known inter-rater error, further complicating interpretation. Additionally, few studies reported whether the interpretation of remote and reference measures was blinded, introducing a potential for bias in outcome assessment. Finally, while the flow and timing domain was largely well addressed, the lack of standardisation in the interval between assessments and test procedures may still contribute to unexplained variability.

This review builds on prior work examining digital ROM measurement methods but uniquely focuses on remote ROM assessment. In contrast to a previous systematic review by Shepherd and colleagues that found no significant differences between digital measurement tools and traditional goniometer tools, our study identifies potential sources of measurement error and highlights specific movement planes and pathological population that are more prone to inaccuracies, thereby highlighting the need for further evaluation and optimisation required in remote assessment contexts. [45]

While this review provides a comprehensive analysis, several limitations must be acknowledged: Heterogeneity across studies in sample populations, measurement techniques, and reference standards contributed to high variability ($I^2 > 90\%$ in meta-analysis). The use of prediction intervals in this review partially addresses this limitation by providing an estimate of the expected range of effects in future settings. Several additional methodological approaches could be used to further explore this heterogeneity. Meta-regression analysis may allow investigation of whether study-level factors, such as population characteristics (e.g. healthy vs pathological), type of technology (e.g. IMU vs camera-based systems), or measurement protocol, contribute to variability in effect estimates.

Additionally, sensitivity analyses could be conducted to assess the robustness of findings by excluding studies with high risk of bias, small sample sizes, or non-standardised measurement approaches. This may help determine whether specific studies disproportionately influence pooled estimates.

However, the feasibility of such analyses in the present review was limited by the relatively small number of studies within each subgroup and variability in reported outcome measures. As such, these approaches may not have yielded reliable or interpretable results. Future research with more standardised reporting and larger datasets would enable more advanced quantitative exploration of heterogeneity.

Limited representation of certain movement planes (e.g., extension) means that findings may not be fully generalisable across all ROM measurements. Potential

publication bias was detected, indicating that negative or low-accuracy findings may be underreported. The demographic profile of the included studies presents an important limitation. Most of the evidence stems from research involving young, healthy individuals, which may not reflect the performance of remote ROM tools in real-world clinical populations—particularly older adults or patients with pathological shoulders, such as rotator cuff tears or adhesive capsulitis. These groups may present with limited mobility, compensatory movement patterns, or pain, all of which can affect measurement accuracy and usability.

The findings of this review suggest that remote technologies may have a role in supporting shoulder assessment, particularly in tele-rehabilitation and follow-up settings, although their variability means they should be used as adjuncts rather than replacements for standard clinical measures. The observed inconsistency in accuracy and reliability highlights the importance of clinician training, particularly in areas such as sensor positioning, patient instruction, and standardisation of measurement protocols to optimise performance. While these technologies may improve access to care and patient engagement, limitations in detecting small but clinically meaningful changes may affect their ability to guide treatment decisions and monitor subtle progress. From a health economic perspective, remote assessment has the potential to reduce healthcare utilisation and associated costs, although this must be balanced against the risk of reduced measurement precision and the potential need for confirmatory in-person assessment. Additionally, future research should address digital inequality, as access to and familiarity with digital technologies may vary by age, socioeconomic status, and health literacy. Ensuring that remote assessment tools are validated across diverse populations will be essential to support effective implementation.

Conclusion

Remote technologies for assessing shoulder range of motion show generally small differences compared to reference standards, but performance is inconsistent, with substantial heterogeneity and variable reliability across studies. Although average error is often below reported minimal clinically important differences (MCID), the wide variability observed suggests these tools may not reliably detect small but clinically meaningful changes. As such, they may have value for tracking overall trends in shoulder function, but their use in clinical decision-making—particularly in pathological populations—should be approached with caution.

References

1. Barzegar Khanghah A, Fernie G, Roshan Fekr A. Joint angle estimation during shoulder abduction exercise using contactless technology. BioMedical Engineering OnLine 2024;23:11. doi:10.1186/s12938-024-01203-5
2. Bechard L, Bell K, Lynch A. Preliminary validation of a mobile force Sensing device for clinical and telerehabilitation. J Biomech 2020;110:109973. doi:10.1016/j.jbiomech.2020.109973
3. Boissy P, Diop-Fallou S, Lebel K, Bernier M, Balg F, Tousignant-Laflamme Y. Trueness and Minimal Detectable Change of Smartphone Inclinometer Measurements of Shoulder Range of Motion. Telemed J E Health 2017;23:503-6. doi:10.1089/tmj.2016.0205
4. Borresen A, Chakka K, Wu R, Lin CK, Wolfe C, Prabhakaran B, Annaswamy TM. Comparison of in-person and synchronous remote musculoskeletal exam using augmented reality and haptics: A pilot study. Pm r 2023;15:891-8. doi:10.1002/pmrj.12883
5. Bradley KE, Cook C, Reinke EK, Vinson EN, Mather RC, III, Riboh J, et al. Comparison of the accuracy of telehealth examination versus clinical examination in the detection of shoulder pathology. Journal of Shoulder and Elbow Surgery 2021;30:1042-52. doi:10.1016/j.jse.2020.08.016
6. Campbell KJ, Louie PK, Bohl DD, Edmiston T, Mikhail C, Li J, et al. A novel, automated text-messaging system is effective in patients undergoing total joint arthroplasty. J Bone Joint Surg Am 2019;101:145-51. doi:10.2106/JBJS.17.01505
7. Chan LYT, Chua CS, Chou SM, Seah RYB, Huang Y, Luo Y, et al. Assessment of shoulder range of motion using a commercially available wearable sensor-a validation study. MHealth 2022;8:30-. doi:10.21037/mhealth-22-7
8. Chen YP, Lin CY, Tsai MJ, Chuang TY, Lee OK. Wearable Motion Sensor Device to Facilitate Rehabilitation in Patients With Shoulder Adhesive Capsulitis: Pilot Study to Assess Feasibility. J Med Internet Res 2020;22:e17032. doi:10.2196/17032
9. Constant CR, Murley AH. A clinical method of functional assessment of the shoulder. Clin Orthop Relat Res 1987:160-4.
10. Çubukçu B, Yüzgeç U, Zileli R, Zileli A. Reliability and validity analyzes of Kinect V2 based measurement system for shoulder motions. Med Eng Phys 2020;76:20-31. doi:10.1016/j.medengphy.2019.10.017
11. Cuesta-Vargas AI, Roldan-Jimenez C. Validity and reliability of arm abduction angle measured on smartphone: a cross-sectional study. BMC Musculoskelet Disord 2016;17:93. doi:10.1186/s12891-016-0957-3
12. Cui J, Yeh SC, Lee SH. Wearable Sensors Integrated with Virtual Reality: A Self-Guided Healthcare System Measuring Shoulder Joint Mobility for Frozen Shoulder. J Healthc Eng 2019;2019:7681237. doi:10.1155/2019/7681237
13. Faber M, Andersen MH, Sevel C, Thorborg K, Bandholm T, Rathleff M. The majority are not performing home-exercises correctly two weeks after their initial instruction-an assessor-blinded study. PeerJ 2015;3:e1102. doi:10.7717/peerj.1102
14. Falbriard M, Meyer F, Mariani B, Millet GP, Aminian K. Drift-Free Foot Orientation Estimation in Running Using Wearable IMU. Front Bioeng Biotechnol 2020;8:65. doi:10.3389/fbioe.2020.00065

15. Gismervik SØ, Drogset JO, Granviken F, Rø M, Leivseth G. Physical examination tests of the shoulder: a systematic review and meta-analysis of diagnostic test performance. BMC Musculoskeletal Disorders 2017;18:41. doi:10.1186/s12891-017-1400-0
16. Gushikem A, Gomes Costa RR, Lima Cabral AL, Lopes Bomtempo LF, de Mendonça Cardoso M. Validity of range of motion, muscle strength, sensitivity, and Tinel sign tele-assessment in adults with traumatic brachial plexus injury. Acta Neurochir (Wien) 2022;164:1317-28. doi:10.1007/s00701-022-05164-3
17. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schünemann HJ. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. Bmj 2008;336:924-6. doi:10.1136/bmj.39489.470347.AD
18. Hayes K, Walton JR, Szomor ZL, Murrell GAC. Reliability of five methods for assessing shoulder range of motion. Australian Journal of Physiotherapy 2001;47:289-94. doi:https://doi.org/10.1016/S0004-9514(14)60274-9
19. Höglund G, Grip H, Öhberg F. The importance of inertial measurement unit placement in assessing upper limb motion. Medical Engineering & Physics 2021;92:1-9. doi:https://doi.org/10.1016/j.medengphy.2021.03.010
20. Howlett N, Trivedi D, Troop NA, Chater AM. Are physical activity interventions for healthy inactive adults effective in promoting behavior change and maintenance, and which behavior change techniques are effective? A systematic review and meta-analysis. Transl Behav Med 2019;9:147-57. doi:10.1093/tbm/iby010
21. Hwang S, Ardebol J, Ghayyad K, Pak T, Bonadiman JA, Denard PJ, Menendez ME. Remote visual estimation of shoulder range of motion has generally high interobserver reliability but limited accuracy. JSES Int 2023;7:2528-33. doi:10.1016/j.jseint.2023.07.002
22. Jayaraman S, Joshy J, Priya PK, Poduval M, Thangavel AB. SmartWatch as a Kinaesthetic System for Shoulder Function Assessment. *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops).* 2020:1-6.
23. Johnson BM, Crawford Z, Harley J, Bonamer JP, Grawe BM. A comparison of quantitative and qualitative shoulder range of motion with patient-reported outcomes. J Shoulder Elbow Surg 2023;32:1364-9. doi:10.1016/j.jse.2023.02.133
24. Kane LT, Thakar O, Jamgochian G, Lazarus MD, Abboud JA, Namdari S, Horneff JG. The role of telehealth as a platform for postoperative visits following rotator cuff repair: a prospective, randomized controlled trial. Journal of Shoulder and Elbow Surgery 2020;29:775-83. doi:10.1016/j.jse.2019.12.004
25. Knapp PW, Keller RA, Mabee KA, Pillai R, Frisch NB. Quantifying Patient Engagement in Total Joint Arthroplasty Using Digital Application-Based Technology. J Arthroplasty 2021;36:3108-17. doi:10.1016/j.arth.2021.04.022
26. Kolber MJ, Hanney WJ. The reliability and concurrent validity of shoulder mobility measurements using a digital inclinometer and goniometer: a technical report. Int J Sports Phys Ther 2012;7:306-13.
27. Lam WWT, Fong KNK. Validity and Reliability of Upper Limb Kinematic Assessment Using a Markerless Motion Capture (MMC) System: A Pilot Study. Arch Phys Med Rehabil 2024;105:673-81.e2. doi:10.1016/j.apmr.2023.10.018
28. Mitchell K, Gutierrez SB, Sutton S, Morton S, Morgenthaler A. Reliability and validity of goniometric iPhone applications for the assessment of active shoulder external

rotation. Physiother Theory Pract 2014;30:521-5. doi:10.3109/09593985.2014.900593
29. Muir SW, Corea CL, Beaupre L. Evaluating change in clinical status: reliability and measures of agreement for the assessment of glenohumeral range of motion. N Am J Sports Phys Ther 2010;5:98-110.
30. Niu X, Zou K, Shen D, Drew S, Wu S, Guo G, Chen R. UltraMotion: High-Precision Ultrasonic Arm Tracking for Real-World Exercises. IEEE Transactions on Mobile Computing 2024;23:1846-62. doi:10.1109/TMC.2023.3241077
31. Ongvisatepaiboon K VV, Chignell M, Mekhora K, Chan JH. Smartphone-based audio-biofeedback system for shoulder joint tele-rehabilitation. J Med Imaging Health Inform 2016;6:1127—34. doi:https://doi.org/10.1166/jmihi.2016.1810.
32. Padua R, de Girolamo L, Grassi A, Cucchi D. Choosing patient-reported outcome measures for shoulder pathology. EFORT Open Rev 2021;6:779-87. doi:10.1302/2058-5241.6.200109
33. Pereira B, Cunha B, Viana P, Lopes M, Melo ASC, Sousa ASP. A Machine Learning App for Monitoring Physical Therapy at Home. Sensors (Basel) 2023;24. doi:10.3390/s24010158
34. Rajkumar A, Vulpi F, Bethi SR, Raghavan P, Kapila V. Usability study of wearable inertial sensors for exergames (WISE) for movement assessment and exercise. Mhealth 2021;7:4. doi:10.21037/mhealth-19-199
35. Ramkumar PN, Haeberle HS, Ramanathan D, Cantrell WA, Navarro SM, Mont MA, et al. Remote patient monitoring using mobile health for total knee arthroplasty: Validation of a wearable and machine learning-based surveillance platform. J Arthroplasty 2019;34:2253-9. doi:10.1016/j.arth.2019.05.021
36. Rangan A, Handoll H, Brealey S, Jefferson L, Keding A, Martin BC, et al. Surgical vs nonsurgical treatment of adults with displaced fractures of the proximal humerus: the PROFHER randomized clinical trial. JAMA 2015;313:1037-47. doi:10.1001/jama.2015.1629
37. Rigoni M, Gill S, Babazadeh S, Elsewaisy O, Gillies H, Nguyen N, et al. Assessment of Shoulder Range of Motion Using a Wireless Inertial Motion Capture Device-A Validation Study. Sensors (Basel) 2019;19. doi:10.3390/s19081781
38. Roldán-Jiménez C, Martin-Martin J, Cuesta-Vargas AI. Reliability of a Smartphone Compared With an Inertial Sensor to Measure Shoulder Mobility: Cross-Sectional Study. JMIR Mhealth Uhealth 2019;7:e13640. doi:10.2196/13640
39. Rong R, Kuo C. Dynamic Soft Tissue Artifacts during Impulsive Loads: Measurement Errors Vary With Wearable Inertial Measurement Unit Sensor Design. IEEE Trans Biomed Eng 2024;71:3275-82. doi:10.1109/tbme.2024.3416378
40. Sahu D, Shah D, Joshi M, Shaikh S, Gaikwad P, Shyam A. Validation of an on-screen application-based measurement of shoulder range of motion over telehealth medium. J Shoulder Elbow Surg 2022;31:201-8. doi:10.1016/j.jse.2021.06.017
41. Sahu D, Shah D, Joshi M, Shaikh S, Gaikwad P, Shyam A. Validation of an on-screen application-based measurement of shoulder range of motion over telehealth medium. J Shoulder Elbow Surg 2022;31:201-8. doi:10.1016/j.jse.2021.06.017
42. Sallis JF, Prochaska JJ, Taylor WC. A review of correlates of physical activity of children and adolescents. Med Sci Sports Exerc 2000;32:963-75. doi:10.1097/00005768-200005000-00014

43. Seo NJ, Fathi MF, Hur P, Crocher V. Modifying Kinect placement to improve upper limb joint angle measurement accuracy. J Hand Ther 2016;29:465-73. doi:10.1016/j.jht.2016.06.010
44. Shah NV, Gold R, Dar QA, Diebo BG, Paulino CB, Naziri Q. Smart Technology and Orthopaedic Surgery: Current Concepts Regarding the Impact of Smartphones and Wearable Technology on Our Patients and Practice. Curr Rev Musculoskelet Med 2021;14:378-91. doi:10.1007/s12178-021-09723-6
45. Shepherd J, Hansjee S, Divall P, Raval P, Singh HP. How do digital range of motion measurement devices 'measure-up' to traditional goniometry in assessing shoulder range of motion? A systematic review and meta-analysis. Shoulder Elbow 2024;16:363-81. doi:10.1177/17585732231195554
46. Shimizu H, Saito T, Kouno C, Shimoura K, Kawabe R, Shinohara Y, et al. Validity and reliability of a smartphone application for self-measurement of active shoulder range of motion in a standing position among healthy adults. JSES Int 2022;6:655-9. doi:10.1016/j.jseint.2022.04.005
47. Soeters R, Damodar D, Borman N, Jacobson K, Shi J, Pillai R, Mehran N. Accuracy of a smartphone software application compared with a handheld goniometer for measuring shoulder range of motion in asymptomatic adults. Orthop J Sports Med 2023;11:23259671231187297-. doi:10.1177/23259671231187297
48. Soeters R, Damodar D, Borman N, Jacobson K, Shi J, Pillai R, Mehran N. Accuracy of a Smartphone Software Application Compared With a Handheld Goniometer for Measuring Shoulder Range of Motion in Asymptomatic Adults. Orthop J Sports Med 2023;11:23259671231187297. doi:10.1177/23259671231187297
49. Sullivan GM. A primer on the validity of assessment instruments. J Grad Med Educ 2011;3:119-20. doi:10.4300/jgme-d-11-00075.1
50. Tozawa R, Ishii N, Onuma R, Kawasaki T. The reliability and validity of joint range of motion measurement using zoom and a smartphone application. J Phys Ther Sci 2023;35:538-41. doi:10.1589/jpts.35.538
51. Wang CH, Hwang YS, Chen YL, Chen CC, Tsai KY. Implementation of interactive games to a shoulder rehabilitation and evaluation system. Technol Health Care 2020;28:431-7. doi:10.3233/thc-202173
52. Wang G, Fiedler AK, Warth RJ, Bailey L, Shupe PG, Gregory JM. Reliability and accuracy of telemedicine-based shoulder examinations. J Shoulder Elbow Surg 2022;31:e369-e75. doi:10.1016/j.jse.2022.04.005
53. Wang XM, Smith DT, Zhu Q. A webcam-based machine learning approach for three-dimensional range of motion evaluation. PLoS One 2023;18:e0293178. doi:10.1371/journal.pone.0293178
54. Werner BC, Holzgrefe RE, Griffin JW, Lyons ML, Cosgrove CT, Hart JM, Brockmeier SF. Validation of an innovative method of shoulder range-of-motion measurement using a smartphone clinometer application. J Shoulder Elbow Surg 2014;23:e275-82. doi:10.1016/j.jse.2014.02.030
55. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155:529-36. doi:10.7326/0003-4819-155-8-201110180-00009

Appendix 1: Full database search strategies

The following search strategies were used for each database. Search terms were adapted as required for database-specific indexing systems (e.g. MeSH terms in MEDLINE) and syntax. The full search strategy for MEDLINE is presented below, with equivalent adaptations applied across the other databases (PubMed, Web of Science, CENTRAL, and IEEE Xplore) using the same combination of keywords and Boolean operators. This approach allows replication of the search process across all databases.

The MEDLINE (Ovid) search strategy was implemented as follows:

Search: Ovid Medline Inception-2024 Searched 10th May 2024	
Topics	Descriptor, keywords and truncations
Remote	1. Remote
	2. Remote Sensing Technology
	3. Remote consultation
	4. Teleworking
	5. Telerehabilitation
	6. Telemedicine
	7. Online Systems
	8. OR/1-7
Shoulder	9. Shoulder
	10. Shoulder pain
	11. Shoulder Impingement Syndrome
	12. Shoulder Joint
	13. Shoulder Fracture*
	14. Shoulder Dislocation
	15. Shoulder Arthroplasty
	16. Shoulder Replacement
	17. Shoulder Prosthesis
	18. Shoulder Injur*
	19. Brachial Plexus Neuritis
	20. Rotator Cuff Tear Arthropathy

	21. Shoulder Patholog*
	22. OR/9-21
Range of motion &	23. Range of motion
Strength	24. Range-of-motion
	25. ROM
	26. Joint flexibility
	27. Kinematic*
	28. Biomechanic*
	29. Strength
	30. Muscle Strength
	31. Dynamomet*
	32. OR/23-31
Accuracy & Reliability	33. Accuracy
	34. Precision
	35. Reliability
	36. Reproducibility
	37. Validity
	38. OR/33-37
Boolean operators	s8. AND s22. AND s32. AND s38.

Final MEDLINE search string:
(Remote OR "Remote Sensing Technology" OR "Remote consultation" OR Teleworking OR Telerehabilitation OR Telemedicine OR "Online Systems") AND (Shoulder OR "Shoulder pain" OR "Shoulder Impingement Syndrome" OR "Shoulder Joint" OR Shoulder Fracture OR Shoulder Dislocation OR Shoulder Arthroplasty OR Shoulder Replacement OR Shoulder Prosthesis OR Shoulder Injur OR "Brachial Plexus Neuritis" OR "Rotator Cuff Tear Arthropathy" OR Shoulder Patholog*) AND ("Range of motion" OR Range-of-motion OR ROM OR "Joint flexibility" OR Kinematic* OR Biomechanic* OR Strength OR "Muscle Strength" OR Dynamomet*) AND (Accuracy OR Precision OR Reliability OR Reproducibility OR Validity)**

For other databases (PubMed, Web of Science, CENTRAL, and IEEE Xplore), equivalent search strategies were used with appropriate adjustments to database-specific syntax and indexing terms. No restrictions were placed on publication date, and all searches were conducted up to May 2024 (subsequently updated to March 2026).

Appendix 2: Signalling questions used for the QUADAS-2 Tool

1. Patient Selection

Risk of Bias Signalling Questions:

- Was a consecutive or random sample of participants enrolled?
- Was a case-control design avoided?
- Did the study avoid inappropriate exclusions (e.g. excluding certain ROM impairments without justification)?

Applicability Concerns:

- Do the included patients reflect the population in which remote ROM assessments would be used (e.g. clinical, post-op, or community settings)?

2. Index Test: Remote ROM Assessment Method

Risk of Bias Signalling Questions:

- Were the index test results interpreted without knowledge of the reference standard results?
- Was the remote ROM assessment conducted and interpreted according to a pre-specified protocol or clearly described procedure?

Applicability Concerns:

- Does the remote assessment method reflect the intended use in practice (e.g. telehealth, unsupervised home setting)?

3. Reference Standard

Risk of Bias Signalling Questions:

- Is the reference standard (e.g. universal goniometer or 3D motion capture) likely to correctly measure the target ROM?
- Were the reference standard results interpreted without knowledge of the index test results?

Applicability Concerns:

- Is the reference standard appropriate and reflective of best clinical practice for shoulder ROM evaluation?analysis

4. Flow and Timing

Risk of Bias Signalling Questions:

- Was there an appropriate time interval between the remote and reference measurements?
- Did all patients receive the same reference standard?
- Were all patients included in the analysis.