

A systematic review of psychometric properties of physical fitness and physical functioning outcome measures applied in adult people with hemophilia

Pierrette Baschung Pfister¹, Andrea Bärlocher^{1,2}, Anne Juanós Solé², Ruud H. Knols^{1,3}, Nadja Pecorelli¹, Eling D. de Bruin^{1,3,4,5}

¹Directory of Research and Education, Physiotherapy-Occupational Research Center, University Hospital Zurich - Switzerland

²Department of Physiotherapy and Occupational Therapy, University Hospital Zurich, Zurich - Switzerland

³Department of Health Sciences and Technology, Institute of Human Movement Sciences and Sport, ETH Zurich, Zurich - Switzerland

⁴OST – Eastern Swiss University of Applied Sciences, Department of Health, St. Gallen - Switzerland

⁵Division of Physiotherapy, Department of Neurobiology, Care Sciences and Society, Karolinska Institutet, Stockholm - Sweden

ABSTRACT

Introduction: Psychometric properties of clinical measurement instruments are crucial because they indicate the quality, reliability, and validity of data collected in clinical settings as well as how accurately and consistently a tool measures change. The aim was to systematically review and report the measurement properties of performance-based (PerFOMS) or clinician-reported (ClinROM) outcome measurement instruments (OMIs) in adult people with hemophilia (PwH). The second aim was to evaluate the methodological quality and the level of evidence of the reported measurement properties.

Methods: Data search, extraction, and evaluation were conducted in accordance with the COSMIN methodology while following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.

Results: The literature search identified 2557 articles. After title and abstract screening, 166 were included for full-text screening. The 14 included studies reported measurement properties of 11 OMI's (5 ClinROM & 6 PerFOMS). Reliability (intra-rater, inter-rater, or test-retest), measurement error, internal consistency, criterion validity, convergent validity, discriminant validity, and responsiveness were reported. The most reported measurement property was convergent validity. The OMI Hemophilia Joint Health Score (HJHS 2.1) fulfilled sufficient methodological quality for convergent validity and moderate Quality of Evidence (QE). None of the other included OMI's reported sufficient measurement properties.

Conclusion: Researchers and clinicians are challenged when comparing studies and selecting appropriate OMI's when evaluating the functioning of adult PwH, due to unsatisfactory reporting of measurement properties, hampered methodological quality, and a low level of evidence. Therefore, studies investigating measurement properties of ClinROM and PerFOMS in adult PwH with rigorous methodology are required.

Keywords: Adult hemophilia, Outcome assessment, Physical fitness, Physical functioning, Psychometric properties, Systematic review

What's already known about this topic?

- For physical fitness and physical functioning outcome measures, there is still limited consensus on the most appropriate measures for tracking long-term physical sustainability, and tools may not effectively distinguish differences at mid-to-high levels of physical capacity in adult people with hemophilia (PwH).

What does the study add?

- This study gives a comprehensive overview of the measurement properties and the quality of evidence of performance-based and clinician-reported outcome measurement instruments (OMIs) assessing physical fitness and physical function in adult PwH.

Received: March 5, 2026

Accepted: September 1, 2026

Published online: October 6, 2026

This article includes supplementary materials

Corresponding author:

Eling D. de Bruin

email: eling.debruin@hest.ethz.ch

Introduction

Hemophilia is a rare bleeding disorder affecting primarily men with an estimated worldwide prevalence of 29.6 cases per 100 000 males (1). It causes prolonged bleeding, especially in hinge joints, leading to joint inflammation, arthropathy, and muscle problems (2). These issues result in pain,

limited range of motion, and weaker muscles, which can promote a sedentary lifestyle. This inactivity further reduces cardiovascular fitness, proprioception, bone health, overall fitness, quality of life, and increases the risk of falls (3-9).

Exercise for adult people with hemophilia (PwH) is safe and has the potential to counteract these health issues, with evidence showing that exercise can improve range of motion, muscle strength, endurance, balance, proprioception, and bone mineralization (2,4,10). A Cochrane review revealed positive effects of exercise without accompanying adverse effects. Nevertheless, the review authors considered their results cautiously because the number of included studies was small, with considerable heterogeneity in outcome measures. This Cochrane report could not pool results and draw definite conclusions regarding the effects of exercise programs in adult PwH (11).

To evaluate the treatment effects of physical exercise programs in adult PwH, assessments with acceptable psychometric measurement properties are required. Otherwise, it cannot be determined whether treatment is (in-) effective, or rather the assessment applied is inadequate for assessing change. The CONsensus-based Standards for the selection of health Measurement Instruments (COSMIN) taxonomy distinguishes three domains of measurement properties: reliability, validity, and responsiveness (12,13).

To date, there is no consensus as to which assessments are most adequate for evaluating changes in physical fitness and physical functions in adult PwH. Physical fitness, defined as “the capacity of an individual to carry out daily tasks with vigor and alertness”, is operationalized as “a set of measurable health and skill-related attributes” such as muscle strength, balance, flexibility, and endurance (14). Physical function is defined as the capacity of an individual to carry out physical activities of daily living such as standing up from a chair or walking (15). The validity, reliability, and responsiveness of assessments that are used in adult PwH have been criticized for being insufficiently described (16). De Kleijn et al. suggested developing a new core set of clinimetric instruments to assess functional health status (17); however, such a set of measures remains to be defined (2).

A first step towards determining recommendable outcome measurement instruments (OMIs) in adult PwH is conducting a systematic review of existing literature to uncover measurement properties of OMIs used.

Several systematic reviews evaluated measurement properties of OMIs in adult PwH (18-23). However, none of them provides an overview of the measurement properties of performance-based and clinician-reported OMIs that evaluate physical fitness and physical function in adult PwH.

The aim of this systematic review is to investigate the measurement properties (reliability, validity and responsiveness) of performance-based and clinician-reported OMIs assessing physical fitness and physical function in adult PwH, as well as to evaluate the methodological quality of the studies and the level of evidence for the measurement properties.

This PROSPERO-registered study (registration number CRD42023311199) was conducted according to the COSMIN methodology and followed the PRISMA guidelines (24-26).

Search strategy

A comprehensive search strategy was developed in cooperation with a professional librarian. Electronic database searches of MEDLINE, EMBASE, CINAHL, Cochrane, PsycINFO, Scopus, PEDRO, and COSMIN were performed between February 25 and April 1, 2022. An update by the librarian was performed in August 2023, and a hand search was performed in November 2025. Key search terms were defined using two individual search filters for the population (hemophilia) and the measurement construct (physical fitness and physical function). These individual filters were combined with an adapted COSMIN search filter for measurement properties recommended by Terwee et al. (27).

Inclusion and exclusion criteria

Included were [1] studies containing hemophilia-specific and generic performance-based outcome measures (Per FOMs) or clinician-reported outcome measures (ClinROMs) aiming to measure physical fitness and physical function (e.g., strength, endurance, range of motion, coordination, balance, postural control, or gait), [2] when the study sample included a mean age of at least 18 years old PwH, [3] in case the aim of the study was to evaluate one or more measurement properties and [4] studies published as an original article in full text written in English or German.

Studies were excluded if they [1] only used PerFOMS or ClinROMs as an OMI (e.g., randomized controlled trials) without evaluating measurement properties, [2] used PerFOMS or ClinROMs to validate other instruments, or [3] only included PROMs or laboratory values.

Study selection

Search results were exported to EndNote 20 (Clarivate, Philadelphia, USA), where duplicate articles were removed. Article titles and abstracts were screened independently by one of three reviewer pairs (AJS&PBP, NP&RHK, PBP&RHK). Disagreements were discussed until a consensus was reached. Full-text articles were reviewed when the title or abstract did not provide sufficient information to discern eligibility. The same reviewer pairs performed full-text screening and discussed differences. If consensus could not be reached, a third reviewer (NP or PBP) was consulted. A manual search of the references in the included articles was performed, and eligible articles were included following the process described above.

Data extraction

Data were extracted by two pairs of reviewers (PBP&AJS, NP&RHK), using a prepared table (25). Extracted information included the country of the performed study, study aim(s), study design and setting, name and version of the evaluated OMI, construct measured by the OMI, characteristics of the raters (profession, professional experience, and training for the OMI), inclusion and exclusion criteria, sample size, patient characteristics (age, type and severity of hemophilia, bleeding history, and medication), the evaluated measurement properties, and the study results.

Process to evaluate measurement properties

The stepwise procedure to evaluate the measurement properties is shown in Figure 1.

Methodological quality (COSMIN Risk of Bias checklist)

The methodological quality of each included study was evaluated by two independent reviewers (AB&PBP) using the COSMIN Risk of Bias (RoB) checklist. The COSMIN RoB checklist was used as a modular tool, and scores were determined by consensus (24,28).

Structural validity, cross-cultural validity, convergent validity, and discriminative validity or known-groups validity were considered as separate measurement properties. The measurement properties of reliability were defined as follows: 1) test-retest reliability if the rater does not play a role, e.g., PROMs, or if there is no information about the rater; 2) inter-rater reliability as reliability between different raters, including the distinction between scoring and administration; 3) intra-rater reliability as reliability based on repeated measurements by the same rater. As PROMS are excluded, criterion box 1 "PROM development" was not used in this systematic review.

Each item of a criteria box was scored as "very good", "adequate", "doubtful", "inadequate", or "not applicable", according to the RoB checklist. The RoB of each measurement property reported in an article was determined as the lowest rating of any standard in that criteria box (the worst-score-count method) (24,25).

Criteria for good measurement properties

The results of measurement properties reported in the selected articles were evaluated according to the criteria of good measurement properties as "sufficient", "insufficient", or "indeterminate" by two reviewers (AB&PB) (29). The final rating was determined by consensus. To determine whether construct validity (convergent validity, discriminative validity) and responsiveness (construct approach) were "sufficient" or "insufficient", the hypotheses regarding correlations between the OMI of interest and the corresponding comparator instruments, as well as hypotheses regarding differences between groups (e.g. healthy control group compared with adult PwH), subgroups (e.g. different severity, age groups), or changes over time (before and after intervention) had to be reported in the selected articles.

The rating "indeterminate" was given for no reported hypothesis.

Summary and quality of evidence

The evidence of the measurement properties was summarized for each OMI separately. If different studies evaluated the same OMI measurement property, results were qualitatively summarized and compared against the criteria for good measurement properties to determine whether the overall measurement property of the OMI was "sufficient", "insufficient", "inconsistent", or "indeterminate". To decide if the measurement properties of construct validity (convergent validity, discriminative validity) and responsiveness (construct approach) were sufficient or not, the review team formulated a set of hypotheses for studies that did not describe a priori hypothesis (24). For convergent validity, the review team developed a grid visualizing the expected correlations between the evaluated constructs of the OMIs of interest and those of the used comparator instruments. The comparator instrument was rated "high quality" if there was reported information about validity and/or responsiveness in any study population.

Separate hypotheses for subscales were only formulated if there was information regarding the measurement properties for the subscales. The interpretation of the correlation coefficients used to formulate the hypothesis was determined as: low or very low (0-0.25), low (0.26-0.49), moderate (0.50-0.69), high (0.70-0.89), very high (0.90-1.00) (30). Where possible, the review team also formulated a hypothesis for discriminative validity. In cases without any described expectations of differences between the groups or subgroups, or if the review team was unable to formulate a good hypothesis because the minimal important change was unknown, the respective results were ignored as proposed in the COSMIN methodology. Hypotheses for responsiveness were formulated analogously (expected correlation with comparator instruments, expected group/subgroup differences, and/or expected changes over time). The study results were then compared against the retrospectively defined hypotheses. If at least 75% of all results were in accordance with the hypothesis, the measurement property was deemed "sufficient". If, within one study, the OMI of interest was compared with different comparator instruments or if different groups/subgroups were compared, the worst-score count method was applied for the RoB of convergent and discriminative validity.

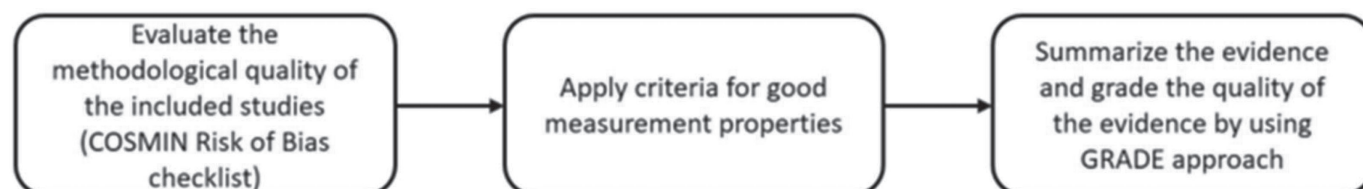


FIGURE 1 - Flow chart search results and screening process.

COSMIN: (COnsensus-based Standards for the selection of health Measurement INstruments) GRADE: Grading of Recommendations, Assessment, Development and Evaluation

Point estimates were considered, and confidence intervals were ignored to rate correlation coefficients for convergent validity and intraclass correlations for reliability. The reason for this is that not all authors presented confidence intervals, and we intended to apply a uniform approach, in addition to avoiding harsher judgment in those instances where confidence intervals were given.

The modified GRADE (Grading of Recommendations, Assessment, Development, and Evaluation) approach was used to define the summarized results. Quality of Evidence (QoE) indicates the extent of confidence that the evidence is trustworthy. QoE is based on [1] overall RoB (quality and number of the studies), [2] inconsistency (of the results of the studies), [3] imprecision (total sample size of the studies), and [4] indirectness (evidence from different populations than the population of interest). QoE was evaluated as “high”, “moderate”, “low”, or “very low”. Reviewers AB&PBP rated QoE, with disagreements resolved by consensus (31). The definition of the quality levels and the downgrading process is shown in Appendix 4. All results, their ratings, and the grading of the evidence were finally summarized in Summary of Findings (SoF) Tables.

Results

The literature search (initial search and update) identified 3335 articles, of which 2557 remained after duplicates were removed. After screening titles and abstracts, 116 articles were included for full-text screening. Fourteen articles met the inclusion criteria and were included in the systematic review (Fig. 2).

The measurement properties of eleven different OMI were investigated in the included articles: five ClinROMS (Colorado Adult Joint Assessment Scale (CAJAS), the Functional Independence Score in Hemophilia (FISH), and three versions of the Hemophilia Joint Health Score (HJHS 1.0, HJHS 1.0_{short}, HJHS 2.1)) and six PerFOMS (Four Square Step Test (FSST), M3 diagnosis (SCHNELL®), Microsoft Kinect V2 sensor, Three-Dimensional Gait Analysis (3DGA), Timed

Up and Go (TUG), and Short Physical Performance Battery (SPPB)). Constructs included: Joint health and gait (32-39), functional independence in self-care, transfers and locomotion (40, 41), voluntary maximal isometric muscle strength (42), and gait and joint function (43), dynamic balance (44), mobility (44), and a combination of lower extremity function, mobility and risk of falling (44). Four articles (33, 34, 37, 45) reported that the included professionals (mostly PTs) were “experienced,” and four articles (34-37) stated that a training session or instruction manual was given.

A summary of the descriptive characteristics of each included study (inclusion and exclusion criteria, sample size, patient characteristics (age, type and severity of hemophilia, bleeding history, and medication), and characteristics of the raters (profession, professional experience, and training for the OMI)) is given in Table 1. A total of 948 adult PwH and 157 healthy controls were included in the review. In seven studies, the age of some patients was lower than 18 years (33-36, 38, 40, 41). However, as these studies also included patients over 18 years of age and reported a mean age of 18 years or above, we included them in our systematic review. Sample sizes were between 10 and 240 individuals, and two articles evaluated the same study sample (32, 39). Table 2 describes the evaluated OMIs and the underlying constructs.

Methodological quality and Measurement Properties

An overview of the COSMIN RoB checklist scores and the measurement properties is presented in Table 3. The most frequently evaluated measurement property is convergent validity, which was mostly rated as “doubtful”. In total, 28 different comparator instruments were used to assess convergent validity. For discriminant validity, values from adult PwH were compared with those from healthy controls, and comparisons of 18 different subgroups were analyzed. The most evaluated subgroups were severity groups. Only one study (37) reported hypotheses regarding correlations between the OMI of interest and the comparator instruments or expected subgroup/group differences. The results were in accordance with the

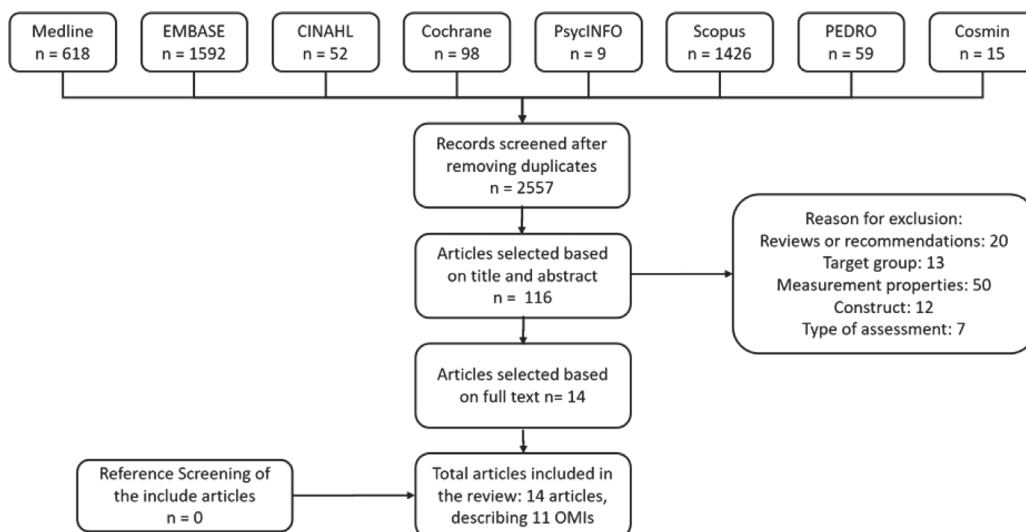


FIGURE 2 - Flow chart search results and screening process. n: number, OMI: Outcome Measurement Instrument

TABLE 1 - Summary of the descriptive characteristics of the included studies

Study (country)	Inclusion and Exclusion Criteria	Sample Size (plus subsample)	Patient characteristics			Characteristics of the raters	
			Age (in years) mean (SD) median (Q1;Q3) Min-Max	Type and Severity of Hemophilia Bleeding History	Medication	Profession / Professional experience	Training for the OMI
Buckner et al. (2017) (32) (USA)	Inclusion criteria Male PwH A or B Age: ≥18 years Any severity of hemophilia With or without inhibitors History of joint pain or joint bleeding Exclusion criteria nr	Total: 381† For HJHS: 240	34 (median) 18–86	Type A: 77% B: 23% Severity Severe: 71% Moderate: 13% Mild: 16% Bleeding History 310 (81%) subjects reported bleeding within the past 6 months (number of bleeds: 7.1 (13.0)).	Inhibitors: 9% Routine infusions: 44% On-demand infusions: 38% Mostly on demand: 19%	PT / nr	nr
Corte- Rodriguez et al. (2020) (33) (Spain)	Inclusion criteria PwH A or B Age: >16 years Any severity of hemophilia Exclusion criteria Acute or subacute bleeding during the study period Values of 0 for all joints on the physical (HJHS 2.1) and ultrasound (HEAD-US) scales Hemarthrosis in the previous 6 weeks. Radiosynovectomy in the past year. History of surgery.	66	34 (14) 16–69	Type A: 85% B: 15% Severity Severe: 80% Moderate: 12% Mild: 8% Bleeding History nr	Prophylaxis: 92% On demand: 8%	Physiatrist/ Extensive experience in managing patients with hemophilia.	nr

(Continued)



TABLE 1 - (Continued)

Study (country)	Inclusion and Exclusion Criteria	Sample Size (plus subsample)	Patient characteristics			Characteristics of the raters	
			Age (in years) mean (SD) median (Q1;Q3) Min-Max	Type and Severity of Hemophilia	Medication	Professional / Professional experience	Training for the OMI
Fischer et al. (2013) (34) (Netherlands)	Inclusion criteria PwH A or B Aged 14 to 30 years Any severity of hemophilia Participating in a skiing trip Exclusion criteria Bleeds or skiing-related injuries	Total: 22 Reliability: 11	Type Total: 20 (17; 24) 14-30 Reliability: 18 (17; 24) Severity Total: Severe: 68% Moderate: 23% Mild: 9% Reliability: Severe: 67% Bleeding History nr	Total: Prophylaxis: 73% Reliability: Prophylaxis: 67%	PT / “Experienced”	All PTs had studied the instruction manual and DVD on how to apply the HJHS and discussed it among themselves before starting the measurements.	
Fischer et al. (2016) (35) (Netherlands and Sweden)	Inclusion criteria PwH A or B Age: >16 years Severe haemophilia Born between 01.01.1970 and 01.01.1994 Lifelong access to care without a history of inhibitors (any inhibitor activity >0.6 BU with decreased recovery) Exclusion criteria nr	Total: 90 NL: 52 SE: 38	Type A: 88% B: 12% Severity Severe: 100% Bleeding History nr	Prophylaxis: 89%	PT / nr	One training session	

Study (country)	Inclusion and Exclusion Criteria	Sample Size (plus subsample)	Patient characteristics				Characteristics of the raters	
			Age (in years) mean (SD) median (Q1;Q3) Min-Max	Type and Severity of Hemophilia Bleeding History	Medication	Professional / Professional experience	Training for the OMI	
Funk et al. (2020) (36) (USA, Romania, Bulgaria, and Argentina)	Inclusion criteria SPINART: PwH A Age: ≥17 years Severe or moderate hemophilia Patients who had completed all patient-reported and clinician-reported outcome assessments	Total: 110 SPINART: 80 Validation study: 30	SPINART: - 31 (mean) - 17-50	Type A: 100% Severity Moderate and severe (numbers not reported)	SPINART: Prophylaxis: 49% On demand: 51% Other: 0	PT / nr		Validity study: - training to administer and score the CAJAS in a live training session - receiving a training manual - video demonstration of the CAJAS patient evaluation of the English version
	Separate validation study: Male PwH A Aged 18 to 50 years Moderate or severe hemophilia (factor VIII: C < 1-2%) Ability to read and understand English. Exclusion criteria SPINART: Very poor joint status, defined as routine need for a wheelchair or inability to walk without assistance of a brace, cane, or crutches, or having three or more ankylosed joints. Separate validation study: Routine need for a wheelchair, two crutches, or canes.		Validation study: - 31 (mean) - 18-47	Bleeding History nr	Validation study: Prophylaxis: 73% On demand: 7% Other: 20%			

(Continued)

TABLE 1 - (Continued)

Study (country)	Inclusion and Exclusion Criteria	Sample Size (plus subsample)	Patient characteristics			Characteristics of the raters	
			Age (in years) mean (SD) median (Q1;Q3) Min-Max	Type and Severity of Hemophilia	Medication	Profession / Professional experience	Training for the OMI
Herbsleb et al. (2010) (42) (Germany)	Inclusion criteria PwH A or B Severe or moderate hemophilia Exclusion criteria nr Healthy control group Exclusion criteria <1 hour of sport/week	Total: 28 Plus 27 healthy controls Reliability: 14 Plus 13 healthy controls	Total: 44 (11) 22-64 Healthy control group: 42 (12) 23-63	Bleeding History	nr	Nr /nr	nr
Lobet et al. (2010) (43) (Belgium)	Inclusion criteria Adult PwH A or B Severe or moderate hemophilia Exclusion criteria Acute joint or muscle bleeding within the last 30 days	18	40 (10) 21-60	Type A: 94% B: 6% Severity Severe: 89% Moderate: 11% Bleeding History nr	Subjects were asked to continue their replacement therapy with clotting factors before and between the 3DGA sessions. Subjects occasionally using NSAID were instructed to observe a wash-out period of at least 72 h prior to testing, whereas those using NSAIDs daily were told not to interrupt the treatment	nr/nr	nr

Study (country)	Inclusion and Exclusion Criteria	Sample Size (plus subsample)	Patient characteristics			Characteristics of the raters	
			Age (in years) mean (SD) median (Q1;Q3) Min-Max	Type and Severity of Hemophilia Bleeding History	Medication	Profession / Professional experience	Training for the OMI
Mateo et al. (2019) (45) (Spain)	Inclusion criteria PwH A or B Age: ≥18 years Severe haemophilia Arthropathy at elbows according to HJHS clinical score	10 Plus 10 healthy controls	37 (10)	Type A: 100%	nr	PT/10 years of experience in haemophilia.	nr
	Severity Severe: 100%						
	Bleeding History nr						
Poonnoose et al. (2005) (40) (India)	Healthy control group 35 (6)	35	21 (mean) 11-40	Type A: 86% B: 24%	nr	nr/nr	nr
	Severity Severe: 100%						
	Bleeding History nr						

(Continued)

TABLE 1 - (Continued)

Study (country)	Inclusion and Exclusion Criteria	Sample Size (plus subsample)	Patient characteristics			Characteristics of the raters	
			Age (in years) mean (SD) median (Q1;Q3) Min-Max	Type and Severity of Hemophilia Bleeding History	Medication	Profession / Professional experience	Training for the OMI
St-Louis et al. (2022) (37) (Canada, Belgium, India, USA, UK)	Inclusion criteria PwH A or B Age: ≥18 years Any severity of hemophilia Exclusion criteria Comorbid illnesses such as rheumatoid arthritis, muscular dystrophy, neurologic disorders, or other non-haemophilia-related significant musculoskeletal conditions that may have independently affected the HJHS score Unable to comply with the study protocol Deemed inappropriate by the investigators. Joint bleed within the last 2 weeks Health control group Inclusion criteria Age-matched healthy men Exclusion criteria Conditions such as rheumatoid arthritis, muscular dystrophy, neurologic disorders, or other significant musculoskeletal conditions that may have independently affected the HJHS	192 Plus 120 healthy controls	35 (median) 18-82	Type A: 85% B: 15% Severity Severe: 68% Moderate: 14% Mild: 18% Bleeding History 40 (21%) subjects reported any bleeds in the 4 weeks before study participation, and 41 (21%) reported joint bleeds in the 12 months before study participation. Healthy control group: 35 (median) 18-69	Prophylaxis: 56% On demand: 44% Nonsteroidal anti-inflammatory drugs: 6% Opioid analgesics: 6% Other analgesics: 12%	PT: Specialized in hemophilia care and has long-standing experience in the use of the HJHS. OT: nr	Standardized set of instructions was provided with the HJHS tool.

Study (country)	Inclusion and Exclusion Criteria	Sample Size (plus subsample)	Patient characteristics			Characteristics of the raters	
			Age (in years) mean (SD) median (Q1;Q3) Min-Max	Type and Severity of Hemophilia Bleeding History	Medication	Profession / Professional experience	Training for the OMI
Tasbihi et al. (2016) (41) (Iran)	Inclusion criteria Male PwH A or B Age: > 10 years Severe hemophilia Exclusion criteria Knee or elbow hemarthrosis that happened in the last 2 weeks. Poor image quality associated with motion artifact or large field-of-view, making it difficult to access small lesions. Not timely referred to the study	25	31 (11) 32 (median) 11-70	Type A: 88% B: 12% Severity Severe: 100% Bleeding History nr	nr	nr/nr	nr
Taylor et al. (2022) (44) (UK)	Inclusion criteria PwH A or B Age: ≥18 years Any severity of hemophilia Registered at the haemophilia centre Able to walk 10 m. Exclusion criteria Unable to stand unaided Any neurological conditions that may affect balance and mobility Non-correctable visual impairment	Total: 80 Inter-rater reliability: 20	Total: 45 (32; 56) Severe: 40 (26;50) Moderate: 42 (37;45) Mild: 52 (37–69) Inter-rater reliability: 38 (27; 49)	Type Total: A: 93% B: 7% Severity Total Severe: 50% Moderate: 13% Mild: 38% Inter-rater reliability: Severe: 85% Moderate: 15% Bleeding History nr	Prophylaxis: 49% On demand: 51% Active inhibitor: 1%	PT / nr	nr

(Continued)

TABLE 1 - (Continued)

Study (country)	Inclusion and Exclusion Criteria	Sample Size (plus subsample)	Patient characteristics			Characteristics of the raters	
			Age (in years) mean (SD) median (Q1;Q3) Min-Max	Type and Severity of Hemophilia Bleeding History	Medication	Profession / Professional experience	Training for the OMI
Uğur et al. (2021) (38) (Turkey)	Inclusion criteria PwH A or B or Von Willebrand disease Exclusion criteria nr	32	21 (median) 15–31	Type A: 72% B: 25% Von Willebrand Disease: 3% Severity Nr Bleeding History nr	Prophylaxis: 91% On demand: 9%	nr /nr	nr
Wang et al. (2017) (39) (USA)	Inclusion criteria Male PwH A or B Age: ≥18 years With or without inhibitors to coagulation factors VIII or IX History of joint pain or bleeding Exclusion criteria nr	Total: 381† For HJHS: 240	34 (median) 18–86	Type A: approximately 75% Severity Severe: 71% Bleeding History 310 subjects reported bleeding within the past 6 months (number of bleeds: 7.1 (13.0))	Inhibitors: 9% Routine infusions: 44% On-demand infusions: 38% Mostly on demand: 19%	nr /nr	nr

†demographic data only from whole sample (n = 381), BU = Bethesda Unit, HEAD-US: Hemophilia Early Arthropathy Detection with Ultrasound, HJHS: Hemophilia Joint Health Score, NL: Netherlands, nr: not reported, PwH: adult People with Hemophilia, SE: Sweden, SPINART†: Trial to Evaluate the Effect of Secondary Prophylaxis With rFVIII Therapy in Severe Hemophilia An Adult and/or Adolescent Subjects Compared to That of Episodic Treatment, UK: United Kingdom, USA: United States of America.

TABLE 2 - Description of the evaluated OMIs

Outcome measurements instrument	Studies	Short description	Measured construct
Colorado Adult Joint Assessment Scale (CAJAS)	Funk et al. (36)	The CAJAS is a ClinROM designed to assess joint health and gait in adult PwH. It comprises nine domains (swelling, muscle atrophy, axial deformity, crepitus, range of motion, contracture, instability, strength, and gait) of six joints (the right and left knee, elbow, and ankle, respectively). Each item is evaluated with a Likert Scale. Items for axial deformity, instability, and gait are assessed using a Likert Scale from 0 to 2; items for swelling, muscle atrophy, crepitus, range of motion, and contracture are assessed from 0 to 3, and strength from 0 to 4. The maximum score for knee and ankle is 25 points; the maximum score for elbow is 21 points. The CAJAS total score ranges between 0 and 142 points, where higher scores indicate more severe joint dysfunction (36).	Joint health and gait
Four Square Step Test (FSST)	Taylor et al. (44)	The FSST is a PerFOM assessing dynamic balance. The test set-up consists of four walking sticks of the same width in a cross formation. Participants step into each square in a clockwise sequence and step back anticlockwise as fast as possible while not being allowed to hit the sticks. The time to complete the task is measured in seconds. Faster times represent better performance, and a result of 15 seconds or more indicates fall risk (44).	Dynamic balance
Functional Independence Score in Hemophilia (FISH) FISH (7 activities)	FISH (7 activities) Poonnoose et al. (40) Tasbihi et al. (41)	The FISH is a ClinROM designed to assess Functional independence in self-care, transfers, and locomotion. The initial version, FISH (7 activities), includes seven items: self-care (eating and grooming, bathing, dressing), transfers (chair, squatting), and locomotion (walking, stairs). Each item can be scored on a Likert Scale from 1 to 4. The total score of FISH (7 activities) ranges between 7 and 28 points. The respective maximum score of 28 means “activity performance without any difficulties like other healthy peers” (40).	Functional independence in self-care, transfers, and locomotion
Hemophilia Joint Health Score (HJHS) HJHS 1.0 HJHS 1.0 _{short} HJHS 2.1	HJHS 1.0 Fischer et al. (35) HJHS 2.1 Fischer et al. (35) St-Louis et al. (37) Taylor et al. (44) Uğur et al. (38) Wang et al. (39)	Hemophilia Joint Health Score (HJHS) is a ClinROM designed to assess joint health and gait for children and adults with hemophilia. There are three different versions: HJHS 1.0, HJHS 1.0 _{short} , HJHS 2.1. The HJHS 1.0 assesses the following items of knees, elbows, and ankles on both sides: range of motion, swelling, duration of swelling, muscle atrophy, axial alignment, crepitus on motion, instability, joint pain, strength, gait at joint level (for knees and ankles only), and global gait. The total score ranges between 0 and 148 points. The HJHS 1.0 _{short} contains the same items except for the global gait. Consequently, the total score can only reach 144 points (34). The HJHS 2.1 includes swelling, duration (of swelling), muscle atrophy, crepitus on motion, flexion loss, extension loss, joint pain, strength, and global gait. The total score ranges between 0 and 124 (33). For all three versions, 0 indicates perfect joint health.	Joint health and gait
M3 diagnos (SCHNELL®)	Herbsleb et al. (42)	M3 diagnosis is a PerFOM used to assess voluntary maximal isometric muscle force of the M. Quadriceps. The maximal voluntary contraction (Fmaxiso) was measured in Newton metre/kilogram (Nm/kg) at 65° knee flexion based on six maximal voluntary muscle contractions for 3-5 seconds each with a three-minute break between the trials. To get familiar with the test, two trials per leg were performed beforehand (42). Higher values indicate more force.	Voluntary maximal isometric muscle force (M. Quadriceps)

(Continued)

TABLE 2 - (Continued)

Outcome measurements instrument	Studies	Short description	Measured construct
Microsoft Kinect V2 sensor (Kinect V2)	Mateo et al. (45)	The Kinect V2 is a ClinROM that is used to measure the range of motion of the elbow. The degree of elbow flexion and extension was measured by a motion capture sensor using an angle-corrected model (45).	ROM of elbow joints
Short Physical Performance Battery (SPPB)	Taylor et al. (2022) (44)	The SPPB is a PerFOM that evaluates lower extremity function. It comprises three components: standing balance, timed three- or four-meter usual-paced walk, and timed tests of five sit-to-stands from a standard chair. Each component is scored from 0 to 4, with a maximum of 12. A higher score indicates better functioning (44).	Lower extremity function
Three-Dimensional Gait Analysis (3DGA)	Lobet et al. (43)	The 3DGA is a PerFOM measuring gait and joint function. It consists of simultaneous analyses of joint kinematics (modeling of body segment movements), kinetics (study of the force interactions of the foot with the ground), and metabolic measurements (calculation of energy and oxygen consumption) by markers on landmarks, six infrared cameras, an oxygen mask, and four 3D strain-gauge force transducers (43). The method focuses on the evaluation of repercussions of segmental abnormalities, and thereby mechanical work, cost, recovery, and efficiency are calculated.	Gait and joint function
Timed Up and Go (TUG):	Taylor et al. (44)	The TUG is a PerFOM that assesses lower extremity function, mobility, and fall risk. In this test, the participant starts in a seated position, with his back against a standard-height chair, without armrests. The participant gets up and walks at a comfortable pace for 10 feet (3 meters), turns around, and returns to a seated position in the chair. The time to perform the task is measured in seconds. Shorter time indicates better mobility. A value of over 13.5 seconds indicates a higher risk of falling in older adults (44).	Lower extremity function, mobility and fall risk

3DGA: three-dimensional gait analysis, BU=Bethesda Unit, CAJAS: Colorado Adult Joint Assessment Scale, FISH: Functional Independence Score in Hemophilia, FSST: Four square step test, HJHS: Hemophilia Joint Health Score, HEAD-US: Hemophilia Early Arthropathy Detection with Ultrasound, SPPB: Short Physical Performance Battery, TUG: Timed Up and Go, UK: United Kingdom, USA: United States of America, 3DGA: Three-dimensional gait analysis.

TABLE 3 - Methodological quality

Measurement Property	OMI	Study	Methodological Quality	Rating
Intra-rater reliability	FSST	Taylor et al, (44)	Doubtful	sufficient
	3DGA	Lobet et al. (43)	Inadequate	sufficient: 25 parameters# insufficient: 7 parameters†
Inter-rater reliability	CAJAS	Funk et al. (36)	Doubtful	(sufficient : Total score and single items
			doubtful	
	FSST	Taylor et al. (44)	Doubtful	sufficient
	HJHS 1.0,	Fischer et al. (34)	Doubtful	sufficient: Total score, ankle right and left, knee right insufficient: Elbow right and left, knee left
	HJHS 1.0 _{short}	Fischer et al. (34)	Doubtful	sufficient: Total score sufficient: Total score, ankle right and left, knee right insufficient: Elbow right and left, knee left
	HJHS 2.1	Fischer et al. (34)	Doubtful	(sufficient: Total score

Measurement Property	OMI	Study	Methodological Quality	Rating
Test-retest reliability	CAJAS	Funk et al. (36)	Doubtful	sufficient: Total score and single joints
	M3 diagnos	Herbsleb et al. (42)	Doubtful	sufficient
Measurement error	FSST	Taylor et al. (44)	Doubtful	indeterminate
	HJHS 1.0	Fischer et al. (34)	Doubtful	indeterminate
	HJHS 1.0 _{short}	Fischer et al. (34)	Doubtful	indeterminate
	HJHS 2.1	Fischer et al. (34)	Doubtful	indeterminate
	M3 diagnos	Herbsleb et al. (42)	Doubtful	indeterminate
	3DGA	Lobet, et al. (43)	Inadequate	indeterminate
Internal consistency	CAJAS	Funk et al. (36)	Inadequate	indeterminate
	Fish (7 activities)	Poonnoose et al. (40)	Inadequate	indeterminate
	HJHS 2.1	St-Louis et al. (37)	Doubtful	indeterminate
		Wang et al. (39)	Doubtful	indeterminate
Criterion validity	Microsoft Kinect V2	Mateo et al. (45)	Very good	indeterminate
Convergent validity	CAJAS	Funk et al. (36)	Adequate - Extended MRI Scale - Haemo-QoL-A - EQ-5D VAS Doubtful - SF-MPQ (pain summary score) Inadequate - Activity list - SF-MPQ: (pain past 4 weeks) - SF-MPQ: (pain current)	indeterminate
	Fish (7 activities)	Poonnoose et al. (40)	Doubtful - WFH (Gilbert) Score - Petterson score - HAQ	indeterminate
		Tasbihi et al. (41)	Adequate - MRI-A scale - MRI-P scale Doubtful - Sonography	indeterminate
Convergent validity	FSST	Taylor et al. (44)	Doubtful - ABC questionnaire: - HAL - HJHS 2.1 - SPPB - TUG Inadequate - Trips	indeterminate

(Continued)

TABLE 3 - (Continued)

Measurement Property	OMI	Study	Methodological Quality	Rating
	HJHS 1.0	Fischer et al. (35)	Doubtful - Petterson Score - HJHS 2.1	indeterminate
	HJHS 1.0 _{short}	Fischer et al. (34)	Adequate - SF36-PF Doubtful - HAL - Pettersson score Inadequate - VAS joint damage	indeterminate
Convergent validity	HJHS 2.1	Corte-Rodriguez et al. (33)	Doubtful - HEAD-US	indeterminate
		Fischer et al. (35)	Adequate - SF36 Doubtful - EQ-5D-5L - HAL - Petterson Score Inadequate - Self-reported bleed	indeterminate
		Jean St-Louis et al. (37)	Doubtful - FISH (8 activities) - HAL - SF-MPQ subscales - WFH (Gilbert) Score)	sufficient
		Uğur et al. (38)	Doubtful - HEAD-US	indeterminate
	m3 diagnos	Herbsleb et al. (42)	Doubtful - VAS pain	indeterminate
Discriminative or known-groups validity	CAJAS	Funk et al. (36)	Doubtful - Age group - Joint damage	indeterminate
	FSST	Taylor et al. (44)	Very good - Severity group	indeterminate
	TUG	Taylor et al. (44)	Very good - Severity group	indeterminate
	SPPB	Taylor et al, (44)	Very good - Severity group	indeterminate

Measurement Property	OMI	Study	Methodological Quality	Rating
Discriminative or known-groups validity	HJHS 2.1	Buckner et al. (32)	Doubtful - presence of anxiety or depression - type of pain - lifetime extent of routine (prophylactic) factor infusion - level of functional impairment - presence of arthritis/ bone/joint problems - current treatment regimen - haemophilia severity	indeterminate
		Jean St-Louis et al. (37)	Doubtful - Severity group Very good - Age group - Healthy controls	sufficient
		Taylor et al. (44)	Very good - Severity group	indeterminate
		Uğur et al. (38)	Doubtful - With or without arthropathy	indeterminate
	m3 diagnos	Herbsleb et al. (42)	Very good - Healthy controls	indeterminate
	Microsoft Kinect V2	Mateo et al. (45)	Very good - Healthy controls	indeterminate
Responsiveness (comparison with other instruments) Responsiveness (comparison between subgroups)	CAJAS	Funk et al. (36)	Doubtful - EQ-5D VAS - Extended MRI Scale - Haemo-QoL-A - SF-MPQ (pain summary score) Inadequate - Activity list - SF-MPQ: (pain past 4 weeks) - SF-MPQ: (pain current	indeterminate
	CAJAS	Funk et al. (36)	Doubtful Three change groups (improvement, no change, deterioration) based on Haemo-QoL-A Physical Functioning and Extended MRI scores.	indeterminate

CAJAS: Colorado Adult Joint Assessment Scale, FISH: Functional Independence Score in Haemophilia, FSST: Four square step test, HJHS: Haemophilia Joint Health Score, SPPB: Short performance physical battery, TUG: Timed up and go, V: Very good 3DGA: three-dimensional gait analysis, # cadence (step/min), step length (metre), stance phase duration (% gait cycle), ankle maximum flexion at loading, response (°), ankle sagittal ROM at push-off phase (°), ankle transversal ROM (°), ankle average transversal position (°), knee flexion at heel strike (°), knee maximum flexion at loading response (°), knee ROM in swing phase (°), hip flexion at heel strike (°), hip sagittal ROM (°), hip frontal ROM (°), ankle max plantarflexion moment at push-off phase (N m kg-1), knee max extension moment at loading response (N m kg-1), hip max extension moment at early loading response (N m kg-1), ankle max power at push-off phase (W kg-1), knee max eccentric power at swing phase (W kg-1), hip max positive power of extensors at early loading response (W kg-1), external work (J kg-1m-1), internal work (J kg-1m-1), total work (J kg-1m-1), recovery (%), cost (J kg-1m-1), efficiency (%), † ankle dorsiflexion at heel strike (°), hip transversal ROM (°), pelvic sagittal ROM (°), pelvic frontal ROM (°), pelvic transversal ROM (°), hip max flexion moment at early swing phase (N m kg-1), hip max positive power of flexors at early swing phase (W kg-1)

hypothesis and therefore rated as “sufficient”. The convergent and discriminative validity of all other studies was—due to the missing hypothesis—rated as “indeterminate”. Reliability was examined nine times, with methodological quality rated as doubtful eight times. The most common reasons were that the authors did not document whether the patients were stable between the repeated measurements and/or did not supply enough information about whether the raters were blinded to the scores or values of other repeated measurement(s) in the same patients. In one case, the methodological quality was rated as inadequate because a time interval of 18 ± 5 weeks (range: 13–33) was deemed to be too long to assess reliability (43). As the minimal important change of the evaluated OMIs was not defined, measurement errors were deemed “indeterminate” for all cases. Internal consistency depends on structural validity, and its QoE cannot be higher than that of structural validity. Because of the lack of information on structural validity for the OMIs of interest, internal consistency was rated as “indeterminate”. One study evaluated criterion validity and had a very high methodological quality, but the rating was “indeterminate” since no correlations with the gold standard were analyzed (45). Responsiveness was only evaluated for CAJAS, and no hypothesis was formulated; consequently, the study had to be rated “indeterminate”. No studies evaluated content, structural, or cross-cultural validity of an evaluated OMI. Also, the criterion approach for responsiveness had not been assessed.

Summary and quality of evidence

The findings are summarized in a Summary of Findings table (SoF table) per OMI (Table 4). Only two OMIs were evaluated in more than one study: HJHS 2.1 was evaluated in four different studies (33, 35, 37, 38) and FISH (7 activities) in two different studies (40, 41). Regarding convergent validity, only St. Louis et al. formulated hypotheses (37). To perform a uniform assessment of all studies, the review team adapted those hypotheses to fit into the developed grid. Thus, all hypotheses mentioned in the SoF tables are formulated by the review team based on the same assumptions. The described QoE is only valid for the specific population described in the last column.

CAJAS, HJHS 1.0 (total score), HJHS1.0_{short} (total score), and M3 diagnosis have “sufficient” reliability and convergent validity, but the QoE of all these measurement properties is “very low”. HJHS 2.1 has also “sufficient” reliability and convergent validity. QoE for convergent validity is moderate, and that of validity is “very low”. FSST has “sufficient” reliability, but the rating for convergent validity was “insufficient”. Twenty-five of 32 parameters of the 3DGA have “sufficient” reliability, and FISH (7 activities) has “sufficient” convergent validity. QoE for convergent validity and reliability of FSST, FISH (7 activities), and 3DGA is “very low”. All other measurement properties (54% of all 35 investigated measurement properties) were rated as “indeterminate”.

The main reason for downgrading the QoE in all investigated measurement properties was overall RoB, which means that the methodological quality and number of the included studies were low, except for the discriminative

validity, which had “no” overall RoB in five studies and was only downgraded once by one point. Inconsistency was “not applicable” (because the measurement properties were only evaluated in one study), or the results were consistent. Imprecision, which is an indication of the sample size, had to be downgraded by 2 points in almost two-thirds of the studies, while “no” imprecision could be found in three studies. Indirectness was only relevant in a few studies.

Discussion

This systematic review included 14 articles evaluating five ClinROMS (CAJAS, FISH with 7 activities, HJHS 1.0, HJHS 1.0_{short}, and HJHS 2.1) and six PerFOMS (FSST, M3 diagnosis (SCHNELL®), Microsoft Kinect V2 sensor, 3DGA, SPPB, and TUG) in adult PwH. The most evaluated measurement property was convergent validity. The OMIs that showed at least moderate quality of evidence were the FSST (discriminative validity), HJHS 2.1 (internal consistency, convergent validity, and discriminative validity), SPPB (discriminative validity), and the TUG (discriminative validity).

Evaluated sample and external validity

The studies in our review were very heterogeneous regarding inclusion criteria, setting, and age. Most studies included adult PwH A or B, but some studies included only people with severe hemophilia or excluded people with mild hemophilia. The lower age limit varied considerably, from > 10 years to ≥ 18 years. We included studies with individuals with an age below 18 years; however, we required a mean sample age of at least 18 years, since the portion of PwH under 18 years could not be distinguished. As Stephenson et al. excluded all studies where data on those aged ≤ 18 years were not clearly distinguished from the adult data, we did not include studies that were already evaluated in the review by Stephenson et al. (46). Most studies included consecutive adult PwH attending the hemophilia comprehensive center or local hospital, but two studies included participants of a special event [downhill skiing trip (34) or hemophilia workshop (38)]. Therefore, the external validity of these results must be considered with caution, and results should be confirmed with studies including all types and severity stages of hemophilia to improve generalizability.

Included OMI and measured constructs

The included OMIs measure physical fitness and physical function. HJHS and CAJAS, primarily measuring the construct of joint health, were also included since attributes of physical fitness such as strength and flexibility are also part of joint health. Three OMIs included in our study have already been examined in previous systematic reviews: FISH and TUG were included in the review of Timmer et al. (23), and HJHS was included in the review of Gouw et al. (19). Both systematic reviews included only articles published before June 2016. Several assessments with variable underlying constructs were included because this has major advantages, such as capturing holistic health and addressing biologic variability of adult PwH (47). Conversely, however, heterogeneity of measurement

TABLE 4 - Summary of findings

Colorado Adult Joint Assessment Scale: CAJAS		
Measurement property	Overall rating	Quality of Evidence (for people with moderate to severe haemophilia, ≥17 years)
Internal consistency	indeterminate	Very low • overall RoB: -3 • Inconsistency: na • Imprecision: -2 • Indirectness: 0
Inter-rater reliability	sufficient	Very low • Overall RoB: -2 • Inconsistency: na • Imprecision: -2 • Indirectness: 0
Test-retest reliability	sufficient	Very low • Overall RoB: -2 • Inconsistency: NA • Imprecision: -2 • Indirectness: 0
Convergent validity	sufficient 7 of 7 (100%) hypothesis were confirmed	Very low • Overall RoB: -2 • Inconsistency: NA • Imprecision: -1 • Indirectness: 0
Responsiveness (comparison with other instruments)	indeterminate	Very low • Overall RoB: -2 • Inconsistency: NA • Imprecision: -1 • Indirectness: 0
Responsiveness (comparison between subgroups)	indeterminate	Very low • Overall RoB: -2 • Inconsistency: na • Imprecision: -1 • Indirectness: 0
Four Square Step Test: FSST		
Measurement property	Overall rating	Quality of Evidence (for adult PwH of all type of severity, ≥18 years)
Intra-rater Reliability	sufficient	Very low • Overall RoB: -2 • Inconsistency: na • Imprecision: -1 • Indirectness: 0
Inter-rater Reliability	sufficient	Very low • Overall RoB: -2 • Inconsistency: na • Imprecision: -2 • Indirectness: 0

(Continued)

TABLE 4 - (Continued)

Measurement error (intra-rater)	indeterminate	Very low - Overall RoB: -2 - Inconsistency: na - Imprecision: -1 - Indirectness: 0
Measurement error (inter-rater)	indeterminate	Very low - Overall RoB: -2 - Inconsistency: na - Imprecision: -2 - Indirectness: 0
Convergent validity	insufficient 3 of 5 (60%) hypothesis were confirmed	Very low - Overall RoB: -2 - Inconsistency: NA - Imprecision: -1 - Indirectness: 0
Discriminative validity	indeterminate	Moderate - Overall RoB: 0 - Inconsistency: NA - Imprecision: -1 - Indirectness: 0
Functional Independence Score in Hemophilia with 7 activities: FISH (7 activities)		
	Overall rating	Quality of Evidence (for people with severe haemophilia, aged >10 years)
Internal consistency	indeterminate	Very low - Overall RoB: -3 - Inconsistency: na - Imprecision: -2 - Indirectness: 0
Convergent validity	sufficient 5 of 6 (83%) hypothesis were confirmed	Very low - Overall RoB: -1 - Inconsistency: 0 - Imprecision: -2 - Indirectness: 0
Haemophilia joint health score, version 1.0: HJHS 1.0		
Measurement property	Overall rating	Quality of Evidence (for PwH of all type of severity aged between 14 and 30 years)
Inter-rater reliability	Total score: sufficient Right/left ankle: sufficient Right knee: sufficient Right/left elbow: insufficient Left knee: insufficient	Very low - Overall RoB: -2 - Inconsistency: NA - Imprecision: -2 - Indirectness: -1 (skiing trip¶)
Measurement error (inter-rater)	indeterminate	Very low - Overall RoB: -2 - Inconsistency: na - Imprecision: -2 - Indirectness: -1 (skiing trip¶)

Convergent validity	sufficient 2 of 2 (100%) hypothesis were confirmed	Very low - Overall RoB: -2 - Inconsistency: na - Imprecision: -2 - Indirectness: -2 (skiing trip¶ low evidence for validity)
Hemophilia joint health score, version 1.0_{short}: HJHS 1.0_{short}		
Measurement property	Overall rating	Quality of Evidence (for PwH of all type of severity aged between 17 and 30 years)
Inter-rater reliability	Total score: sufficient Right knee: sufficient Right/left ankle: sufficient Right/left elbow: insufficient Left knee: insufficient	Very low - Overall RoB: -2 - Inconsistency: NA - Imprecision: -2 - Indirectness: -1 (skiing trip¶)
Measurement error (inter-rater)	indeterminate	Very low - Overall RoB: -2 - Inconsistency: NA - Imprecision: -2 - Indirectness: -1 (skiing trip¶)
Convergent validity	sufficient 2 of 3 (75%) hypothesis were confirmed	Very low - Overall RoB: -2 - Inconsistency: NA - Imprecision: -2 - Indirectness: -1 (skiing trip¶)
Hemophilia joint health score, version 2.1: HJHS 2.1		
Measurement property	Overall rating	Quality of Evidence (for PwH of all type of severity, aged ≥14 years‡)
Internal consistency	indeterminate	Moderate ^b - Overall RoB: -1 - Inconsistency: 0 - Imprecision: 0 - Indirectness: 0
Inter-rater reliability	sufficient	Very low - Overall RoB: -2 - Inconsistency: NA - Imprecision: -2 - Indirectness: -1 (skiing trip¶)
Measurement error (inter-rater)	indeterminate	Very low - Overall RoB: -2 - Inconsistency: na - Imprecision: -2 - Indirectness: -1 (skiing trip¶)
Convergent validity	sufficient 15 of 16 (94%) hypothesis were confirmed	Moderate ^b - Overall RoB: -1 - Inconsistency: 0 - Imprecision: 0 - Indirectness: 0

(Continued)

TABLE 4 - (Continued)

Discriminative validity	indeterminate	Moderate - Overall RoB: -1 - Inconsistency: 0 - Imprecision: 0 - Indirectness: 0
M3 diagnos (SCHNELL®)		
Measurement property	Overall rating	Quality of Evidence (for people with moderate to severe haemophilia, aged ≥22 years)
Test-retest reliability	sufficient	Very low - Overall RoB: -2 - Inconsistency: NA - Imprecision: -2 - Indirectness: 0
Measurement error (test-retest)	indeterminate	Very low - Overall RoB: -2 - Inconsistency: NA - Imprecision: -2 - Indirectness: 0
Convergent validity	sufficient 2 of 2 (100%) hypothesis were confirmed	Very low - Overall RoB: -2 - Inconsistency: NA - Imprecision: -2 - Indirectness: -1 (low evidence for validity)
Discriminative validity	indeterminate	Low - Overall RoB: 0 - Inconsistency: NA - Imprecision: -2 - Indirectness: 0
Microsoft Kinect V2 Sensor		
Measurement property	Overall rating	Quality of Evidence (for people with severe hemophilia, aged ≥18 years)
Criterion validity	indeterminate	Low - Overall RoB: 0 - Inconsistency: NA - Imprecision: -2 - Indirectness: 0
discriminative validity	indeterminate	Low - Overall RoB: 0 - Inconsistency: NA - Imprecision: -2 - Indirectness: 0
Short performance physical battery: SPPB		
Measurement property	Overall rating	Quality of Evidence (for PwH of all type of severity, aged ≥18 years)
Discriminative validity	indeterminate	Moderate - Overall RoB: 0 - Inconsistency: NA - Imprecision: -1 - Indirectness: 0

Three-dimensional gait analysis: 3DGA		
Measurement property	Overall rating	Quality of Evidence (for people with moderate to severe haemophilia, aged ≥21 years)
Intra-rater reliability	25 of 32 parameters: sufficient	Very low - Overall RoB: -3 - Inconsistency: NA - Imprecision: -2 - Indirectness: 0
Measurement error (intra-rater)	indeterminate	Very low - Overall RoB: -3 - Inconsistency: NA - Imprecision: -2 - Indirectness: 0
Timed up and go: TUG		
Measurement property	Overall rating	Quality of Evidence (for PwH of all type of severity, aged ≥18 years)
Discriminative validity	indeterminate	Moderate - Overall RoB: 0 - Inconsistency: NA - Imprecision: -1 - Indirectness: 0

¶PwH were recruited during a skiing trip, †Wang et al. (2017 and Fischer et al. (2016) included only people with severe hemophilia, §The lower age limit of the included PwH ranged between 14 and 18 years, dependent on the study (see Table 7), na: not applicable, RoB: Risk of Bias

instruments may challenge metric standardization and enhance methodological bias. Heterogeneity of assessments may be improved by establishing standardized core outcome sets for adult PwH; e.g., consensus-driven frameworks in which researchers in the area of PwH research agree on a minimum set of specific outcomes that should be measured and reported in clinical trials for this specific condition.

Timmer et al. revealed moderate QoE for hypothesis testing (convergent/discriminative validity) and limited evidence for reliability of the FISH (23). In contrast to their review, we did not adjust the COSMIN criteria for minimum sample size to 20 and therefore rated the QoE of all evaluated measurement properties of the FISH as “very low”. They included three articles (48-50) that reported on the TUG; none of them aimed to evaluate measurement properties. The only study assessing measurement properties of the TUG was published after 2016 (44).

Gouw et al. (19) concluded that the HJHS is the most extensively studied OMI to assess joint health and yielded moderate QoE for hypothesis testing (convergent and discriminative validity) in adult PwH. Although six articles evaluating the HJHS in 610 adult PwHs were published since their review, the level of evidence has not changed. The reason is that the RoB of all additional studies on convergent validity was “doubtful”. Consequently, overall RoB was downgraded by one level for HJHS 2.1 and by two levels for HJHS 1.0 and HJHS 1.0_{short}, leading to moderate or very low QoE, respectively. As no study assessed structural validity following the COSMIN guidelines, we rated structural validity and internal consistency as “indeterminate,” and as the minimal

important change of the HJHS is not known, we rated discriminative validity as “indeterminate.”

In accordance with our review, Stephensen et al. (46) identified the HJHS as the best-evaluated OMI in children with promising measurement properties, as we identified the HJHS 2.1 as the only OMI with a measurement property (convergent validity) with “moderate” QoE. The HJHS is the most widely used clinical tool to evaluate joint health and its impact on physical function in adult PwH. The HJHS 2.1 is the current standard. Since certain measurement properties of the older versions (HJHS 1.0 full and short version) were also examined, we have included these as well.

We included 11 OMIs measuring eight different constructs for adult PwH. Still, relevant aspects of physical fitness and physical function are not covered by those OMIs. Although aerobic capacity is crucial in adult PwH (51), no study could be found that included an OMI evaluating cardiovascular endurance, while Stephensen et al. found OMIs measuring maximal aerobic capacity in children with hemophilia (46). Zetterberg et al. (4) listed several additional constructs measured in hemophilia that are not covered in our review, such as endurance, proprioception, as well as joint range of motion and muscle strength for the mostly affected joints in hemophilia [knee, elbow, ankle, hip, and shoulder (52)]. The presently available OMIs for adult PwH do not reflect all elements of physical fitness and physical function, demonstrating a clear scientific gap.

For optimal use in daily clinical practice, point-of-care tests are most helpful and can save time and money (53). Three

OMIs of this review (m3 diagnosis, 3DGA, Microsoft Kinect V2 sensor) need specific, high-cost equipment. This might be relevant for clinical studies. But for clinical use, there is a considerable need for more evidence for point-of-care tests. A recent study identified the following assessments of performance-based physical function that are practical in the clinical setting: the one-leg balance, tandem stance, 6-min walk test, timed up and down stairs, 30-s sit-to-stand, and timed up and go test (54). Our review reveals that none of these assessments has been sufficiently examined regarding their measurement properties in adult PwH.

Depending on the aim of the study, hemophilia-specific or generic instruments should be used. Hemophilia-specific instruments capture more micro-level data such as joint damage and are more responsive than generic instruments (55). On the other hand, generic tools are more suitable for comparisons with healthy controls or other patients with other, more common chronic diseases (56). Given these different objectives, it can also be assumed that their psychometric properties will differ.

Methodological quality and Measurement Properties

Terminologies in the field of measurement properties are very diverse, and the COSMIN methodology and definitions were only published in 2018 and 2020, respectively (24, 25). Especially for validity and responsiveness, different terms and definitions were used. The concept of reliability was used more homogeneously: although the terminology differed, the meaning was the same (relative reliability was used for consistency of measurement and evaluated with ICC, and absolute reliability was used for measuring standard error of measurement) (57). In this review, we adhered strictly to the COSMIN methodology. Measurement properties were specified according to COSMIN and, if necessary, renamed.

The most frequently evaluated measurement property is convergent validity, which was mostly rated as “doubtful”. This is explained by the fact that the methodological quality depends on the measurement properties of the comparator instruments. Since only a few of the comparator instruments had been evaluated previously in adult PwH, only a few studies reached the methodological quality “adequate”. In addition, very often QoE had to be downgraded by one or two levels due to the small sample size. However, four studies (32, 36, 37, 39) were able to include the recommended minimum of 100 participants.

In this systematic review, only one study evaluated criterion validity using the goniometer as the gold standard. Even though the reliability of the goniometer has been questioned, it is a commonly used instrument for measuring ROM in clinical practice and research and is often used as a gold standard (58-60). For many constructs such as joint health, joint function, or lower extremity function, a gold standard is missing, and criterion validity can therefore not be evaluated (12).

One important finding of the review is that the rating of internal consistency, measurement error, discriminative validity, and responsiveness had to be rated as “indeterminate” in all included OMIs. Although this rating has no relevance and cannot be interpreted, we included the results in the SoF tables. The reason is that this rating depends on

other findings, such as structural validity and minimal important change (31). In other words, once this missing finding is known, the QoE can be derived using our SoF table.

Summary and quality of evidence

Quality of evidence (QoE) was predominantly very low or low (85%), with no high QoE ratings; only a few properties reached moderate QoE (15%). The HJHS 2.1 is the only OMI in adult PwH with one measurement property rated as sufficient (namely convergent validity) and at least moderate QoE, and may be recommended for use in research and clinical practice. All other OMIs have either “insufficient” or “indeterminate” ratings for the criteria for good measurement properties or “low” or “very low” QoE in this population. These results emphasize the need for more rigorous studies to validate existing OMIs measuring physical fitness and physical function in adult PwH. Current OMIs should be used cautiously, as their measurement properties and evidence supporting their use are generally weak, which could impact the validity of clinical outcomes and research findings.

Strengths and limitations of the review

Strengths of the current study are that our search strategy was open, and we did not use a preselection of OMIs by expert opinion. Independent analysis of the studies included by two investigators was given at all phases, and the COSMIN guideline was strictly followed. The scientific gaps are clearly displayed in the SoF tables. New evidence gained can simply be added, facilitating future reporting of an updated QoE of the included OMIs.

Our study also has some limitations. First, as we focused on measurement properties of physical fitness and physical function, feasibility and interpretability were neither evaluated nor discussed. Second, we also evaluated studies that included PwH who were younger than 18 years, as long as the mean of the included sample was at least 18 years. These different age cutoffs could bias our results. Third, we excluded PROMs from our review. Lastly, the studies included were heterogeneous; their methodological quality was mainly poor, and not all measurement properties were evaluated. In particular, content validity, which is supposed to be the most important measurement property, was never addressed (61). If a measure fails to assess what it intends to capture, evaluating other properties is meaningless. Content validity should therefore be evaluated in future studies.

Conclusion

Overall, through this review we were able to critically appraise, compare and summarize the measurement properties of 11 assessments reported in scientific literature that measure physical fitness and physical function in adult PwH. None of the included OMIs have sufficiently reported measurement properties, and the quality of evidence is mostly low or very low. The HJHS 2.1 is the most promising OMI with moderate quality of evidence for sufficient convergent validity. Due to the not satisfactory quality of the available studies and the retrieved QoE of the OMIs, and regarding the

rating “indeterminate” for the criteria for good measurement properties in 54% of the studies, it remains challenging for clinicians to select appropriate instruments when assessing adult PwH. For researchers, it is difficult to compare studies with each other.

There is an identified need for reliable, valid, and responsive OMs to evaluate physical fitness and physical function in adult PWH. Therefore, future studies should evaluate the measurement properties of OMs such as FSST, SPPB, and TUG. Development of a reliable, valid, and responsive set of tests will allow future determination of the effectiveness of training programs which, in turn, may enhance the credibility and accountability of clinical interventions. We recommend following COSMIN reporting guidelines (62) when conducting future studies evaluating assessments in adult PWH.

Acknowledgments

The authors thank Dr. Martina Gosteli (Librarian at Zurich University) for performing the search strategy. Further thanks go to Dr. Anna Grynfeld Smith for proofreading the manuscript for English and structure. NOVO Nordisk Pharma AG, Zurich, Switzerland, is cordially acknowledged for their independent research grant.

Disclosures

Conflict of interest: No potential conflict of interest was reported by the authors during any stage of this systematic review.

Financial support: The authors received a free research Grant for this systematic review from NOVO NORDISK PHARMA, Zurich, Switzerland.

Data Availability Statement: The data presented in this study are available as supplementary material to this article upon reasonable request.

Clinical Trial Protocol number: This study was registered at the International Prospective Register of Systematic Reviews (PROSPERO, registration number CRD4202331199).

References

- lorio A, Stonebraker JS, Chambost H, et al.; Data and Demographics Committee of the World Federation of Hemophilia. Establishing the prevalence and prevalence at birth of hemophilia in males: a meta-analytic approach using national registries. *Ann Intern Med.* 2019;171(8):540-546. [CrossRef PubMed](#)
- Srivastava A, Santagostino E, Dougall A, et al. WFH guidelines for the management of hemophilia, 3rd edition. *Haemophilia.* 2020; 26(Suppl 6): 1-158. [CrossRef](#)
- Groen WG, Takken T, van der Net J, et al. Habitual physical activity in Dutch children and adolescents with haemophilia. *Haemophilia.* 2011;17(5):e906-e912. [CrossRef PubMed](#)
- Zetterberg E, Ljungkvist M, Salim M. Impact of exercise on hemophilia. *Semin Thromb Hemost.* 2018;44(8):787-795. [CrossRef PubMed](#)
- Negrier C, Seuser A, Forsyth A, et al. The benefits of exercise for patients with haemophilia and recommendations for safe and effective physical activity. *Haemophilia.* 2013;19(4):487-498. [CrossRef PubMed](#)
- Forsyth AL, Quon DV, Konkole BA. Role of exercise and physical activity on haemophilic arthropathy, fall prevention and osteoporosis. *Haemophilia.* 2011;17(5):e870-e876. [CrossRef PubMed](#)
- Wang M, Álvarez-Román MT, Chowdary P, et al. Physical activity in individuals with haemophilia and experience with recombinant factor VIII Fc fusion protein and recombinant factor IX Fc fusion protein for the treatment of active patients: a literature review and case reports. *Blood Coagul Fibrinolysis.* 2016;27(7):737-744. [CrossRef PubMed](#)
- Rehm H, Schmolders J, Koob S, et al. Falling and fall risk in adult patients with severe haemophilia. *Hämostaseologie.* 2017;37(2):97-103. [CrossRef PubMed](#)
- Boycott KM, Rath A, Chong JX, et al. International cooperation to enable the diagnosis of all rare genetic diseases. *Am J Hum Genet.* 2017;100(5):695-705. [CrossRef PubMed](#)
- Wagner B, Krüger S, Hilberg T, et al. The effect of resistance exercise on strength and safety outcome for people with haemophilia: a systematic review. *Haemophilia.* 2020;26(2):200-215. [CrossRef PubMed](#)
- Strike K, Mulder K, Michael R. Exercise for haemophilia. *Cochrane Database Syst Rev.* 2016;12(12):CD011180. [PubMed](#)
- de Vet HCW, Terwee CB, Mokkink LB, et al. Measurement in medicine: a practical guide. Practical guides to biostatistics and epidemiology. 2011: Cambridge : Cambridge University Press. [CrossRef](#)
- Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Qual Life Res.* 2010;19(4):539-549. [CrossRef PubMed](#)
- Caspersen CJ, Powell KE, Christenson GM. Physical activity, exercise, and physical fitness: definitions and distinctions for health-related research. *Public Health Rep.* 1985;100(2):126-131. [PubMed](#)
- Garber CE, Blissmer B, Deschenes MR, et al.; American College of Sports Medicine. American College of Sports Medicine position stand. Quantity and quality of exercise for developing and maintaining cardiorespiratory, musculoskeletal, and neuromotor fitness in apparently healthy adults: guidance for prescribing exercise. *Med Sci Sports Exerc.* 2011;43(7):1334-1359. [CrossRef PubMed](#)
- Stephensen D, Bladen M, McLaughlin P. Recent advances in musculoskeletal physiotherapy for haemophilia. *Ther Adv Hematol.* 2018;9(8):227-237. [CrossRef PubMed](#)
- De Kleijn P, Heijnen L, Van Meeteren NL. Clinimetric instruments to assess functional health status in patients with haemophilia: a literature review. *Haemophilia.* 2002;8(3):419-427. [CrossRef PubMed](#)
- Boehlen F, Graf L, Berntorp E. Outcome measures in haemophilia: a systematic review. *Eur J Haematol Suppl.* 2014;76(s76): 2-15. [CrossRef PubMed](#)
- Gouw SC, Timmer MA, Srivastava A, et al. Measurement of joint health in persons with haemophilia: a systematic review of the measurement properties of haemophilia-specific instruments. *Haemophilia.* 2019;25(1):e1-e10. [CrossRef PubMed](#)
- Holdsworth C, Bladen M, Harbidge H, et al. Measuring physical function capacity in persons with haemophilia: a systematic review of performance-based methods. *Haemophilia.* 2025;31(5):840-864. [CrossRef PubMed](#)
- Limberg PF, Terwee CB, Young NL, et al. Health-related quality of life questionnaires in individuals with haemophilia: a systematic review of their measurement properties. *Haemophilia.* 2017;23(4):497-510. [CrossRef PubMed](#)
- Szende A, Schramm W, Flood E, et al. Health-related quality of life assessment in adult haemophilia patients: a systematic review and evaluation of instruments. *Haemophilia.* 2003;9(6): 678-687. [CrossRef PubMed](#)

23. Timmer MA, Gouw SC, Feldman BM, et al. Measuring activities and participation in persons with haemophilia: a systematic review of commonly used instruments. *Haemophilia*. 2018;24(2):e33-e49. [CrossRef PubMed](#)
24. Mokkink LB, de Vet HCW, Prinsen CAC, et al. COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. *Qual Life Res*. 2018;27(5):1171-1179. [CrossRef PubMed](#)
25. Mokkink LB, Boers M, van der Vleuten CPM, et al. COSMIN Risk of Bias tool to assess the quality of studies on reliability or measurement error of outcome measurement instruments: a Delphi study. *BMC Med Res Methodol*. 2020;20(1):293. [CrossRef PubMed](#)
26. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71. [CrossRef PubMed](#)
27. Terwee CB, Jansma EP, Riphagen II, et al. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. *Qual Life Res*. 2009;18(8):1115-1123. [CrossRef PubMed](#)
28. Elsmann EBM, Mokkink LB, Terwee CB, et al. Guideline for reporting systematic reviews of outcome measurement instruments (OMIs): PRISMA-COSMIN for OMIs 2024. *J Clin Epidemiol*. 2024;173:111422. [CrossRef PubMed](#)
29. Terwee CB, Mokkink LB, Knol DL, et al. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Qual Life Res*. 2012;21(4):651-657. [CrossRef PubMed](#)
30. Munro BH. *Statistical methods for health care research*. 4th ed. Lippincott; 2001.
31. Prinsen CAC, Mokkink LB, Bouter LM, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual Life Res*. 2018;27(5):1147-1157. [CrossRef PubMed](#)
32. Buckner TW, Wang M, Cooper DL, et al. Known-group validity of patient-reported outcome instruments and hemophilia joint health score v2.1 in US adults with hemophilia: results from the Pain, Functional Impairment, and Quality of Life (P-FiQ) study. *Patient Prefer Adherence*. 2017;11:1745-1753. [CrossRef PubMed](#)
33. De la Corte-Rodriguez H, Rodriguez-Merchan EC, Alvarez-Roman MT, et al. HJHS 2.1 and HEAD-US assessment in the hemophilic joints: how do their findings compare? *Blood Coagul Fibrinolysis*. 2020;31(6):387-392. [CrossRef PubMed](#)
34. Fischer K, de Kleijn P. Using the Haemophilia Joint Health Score for assessment of teenagers and young adults: exploring reliability and validity. *Haemophilia*. 2013;19(6):944-950. [CrossRef PubMed](#)
35. Fischer K, Nijdam A, Holmström M, et al. Evaluating outcome of prophylaxis in haemophilia: objective and self-reported instruments should be combined. *Haemophilia*. 2016;22(2):e80-e86. [CrossRef PubMed](#)
36. Funk SM, Engelen S, Benjamin K, et al. Validity and reliability of the Colorado Adult Joint Assessment Scale in adults with moderate-severe hemophilia A. *J Thromb Haemost*. 2020;18(2):285-294. [CrossRef PubMed](#)
37. St-Louis J, Abad A, Funk S, et al. The Hemophilia Joint Health Score version 2.1 Validation in Adult Patients study: a multicenter international study. *Res Pract Thromb Haemost*. 2022;6(2):e12690. [CrossRef PubMed](#)
38. Uğur MC, Tamsel İ, Tat NM, et al. Investigation of the relationship between hemophilia joint health score and haemophilia early arthropathy detection with ultrasound score in hemophilic arthropathy. *Türk Klin Tip Bilim Derg*. 2021;41(1):46-50. [CrossRef](#)
39. Wang M, Batt K, Kessler C, et al. Internal consistency and item-total correlation of patient-reported outcome instruments and hemophilia joint health score v2.1 in US adult people with hemophilia: results from the Pain, Functional Impairment, and Quality of Life (P-FiQ) study. *Patient Prefer Adherence*. 2017;11:1831-1839. [CrossRef PubMed](#)
40. Poonnoose PM, Manigandan C, Thomas R, et al. Functional Independence Score in Haemophilia: a new performance-based instrument to measure disability. *Haemophilia*. 2005;11(6):598-602. [CrossRef PubMed](#)
41. Tasbihi M, Pishdad P, Haghpanah S, et al. A comparison between MRI, sonography and Functional Independence Score in Haemophilia methods in diagnosis, evaluation and classification of arthropathy in severe haemophilia A and B. *Blood Coagul Fibrinolysis*. 2016;27(2):131-135. [CrossRef PubMed](#)
42. Herbsleb M, Tutzschke R, Czepa D, et al. Maximal isometric strength measures of the quadriceps muscles. Feasibility and reliability in patients with haemophilia. *Hämostaseologie*. 2010;30(suppl 1):S97-S103. [PubMed](#)
43. Lobet S, Detrembleur C, Francq B, et al. Natural progression of blood-induced joint damage in patients with haemophilia: clinical relevance and reproducibility of three-dimensional gait analysis. *Haemophilia*. 2010;16(5):813-821. [CrossRef PubMed](#)
44. Taylor S, Pemberton S, Barker K. Validity of the four-square step test in persons with haemophilia. *Haemophilia*. 2022;28(2):334-342. [CrossRef PubMed](#)
45. Mateo F, Carrasco JJ, Aguilar-Rodríguez M, et al. Assessment of Kinect V2 for elbow range of motion estimation in people with haemophilia using an angle correction model. *Haemophilia*. 2019;25(3):e165-e173. [CrossRef PubMed](#)
46. Stephensen D, Drechsler W, Scott OM. Outcome measures monitoring physical function in children with haemophilia: a systematic review. *Haemophilia*. 2014;20(3):306-321. [CrossRef PubMed](#)
47. Chmelo EA, Crotts CI, Newman JC, et al. Heterogeneity of physical function responses to exercise training in older adults. *J Am Geriatr Soc*. 2015;63(3):462-469. [CrossRef PubMed](#)
48. Baumgardner J, Elon L, Antun A, et al. Physical activity and functional abilities in adult males with haemophilia: a cross-sectional survey from a single US haemophilia treatment centre. *Haemophilia*. 2013;19(4):551-557. [CrossRef PubMed](#)
49. Fearn M, Hill K, Williams S, et al. Balance dysfunction in adults with haemophilia. *Haemophilia*. 2010;16(4):606-614. [CrossRef PubMed](#)
50. van Genderen FR, Westers P, Heijnen L, et al. Measuring patients' perceptions on their functional abilities: validation of the Haemophilia Activities List. *Haemophilia*. 2006;12(1):36-46. [CrossRef PubMed](#)
51. Tomschi F, Ransmann P, Hilberg T. Aerobic exercise in patients with haemophilia: a systematic review on safety, feasibility and health effects. *Haemophilia*. 2022;28(3):397-408. [CrossRef PubMed](#)
52. Luck JV Jr, Silva M, Rodriguez-Merchan EC, et al. Hemophilic arthropathy. *J Am Acad Orthop Surg*. 2004;12(4):234-245. [CrossRef PubMed](#)
53. Sumita NM, Ferreira CES, Martino MDV, et al. Clinical applications of point-of-care testing in different conditions. *Clin Lab*. 2018;64(7):1105-1112. [CrossRef PubMed](#)
54. Bladen M, Harbidge H, Drechsler W, et al. Identifying performance-based outcome measures of physical function in people with haemophilia (IPOP). *Haemophilia*. 2023;29(6):1611-1620. [CrossRef PubMed](#)
55. Wiebe S, Guyatt G, Weaver B, et al. Comparative responsiveness of generic and specific quality-of-life instruments. *J Clin Epidemiol*. 2003;56(1):52-60. [CrossRef PubMed](#)
56. Klamroth R, Pollmann H, Hermans C, et al. The relative burden of haemophilia A and the impact of target joint development

- on health-related quality of life: results from the ADVATE Post-Authorization Safety Surveillance (PASS) study. *Haemophilia*. 2011;17(3):412-421. [CrossRef PubMed](#)
57. Overend T, Anderson C, Sawant A, et al. Relative and absolute reliability of physical function measures in people with end-stage renal disease. *Physiother Can*. 2010;62(2):122-128. [CrossRef PubMed](#)
58. Li MK, Robinson PM, van Rensburg L. Accuracy of patient-reported range of elbow motion. *Shoulder Elbow*. 2016;8(2): 118-123. [CrossRef PubMed](#)
59. Horsch A, Kleiber S, Ghandour M, et al. Validation of a new equinometer device for measuring ankle range of motion in patients with cerebral palsy: an observational study. *Medicine (Baltimore)*. 2022;101(17):e29230. [CrossRef PubMed](#)
60. Chapleau J, Canet F, Petit Y, et al. Validity of goniometric elbow measurements: comparative study with a radiographic method. *Clin Orthop Relat Res*. 2011;469(11):3134-3140. [CrossRef PubMed](#)
61. Terwee CB, Prinsen CAC, Chiarotto A, et al. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. *Qual Life Res*. 2018;27(5):1159-1170. [CrossRef PubMed](#)
62. Gagnier JJ, Lai J, Mokkink LB, et al. COSMIN reporting guideline for studies on measurement properties of patient-reported outcome measures. *Qual Life Res*. 2021;30(8):2197-2218. [CrossRef PubMed](#)