BMJ Open Diabetes Research & Care

Systematic review on the measurement properties of diabetes-specific patient-reported outcome measures (PROMs) for measuring physical functioning in people with type 2 diabetes

Ellen B M Elsman ,[1] Lidwine B Mokkink,[1,2] Marlous Langendoen-Gort,[3] Femke Rutters,[1,2] Joline Beulens ,[1] Petra J M Elders,[2,3] Caroline B Terwee[1,2]

To cite: Elsman EBM, Mokkink LB, Langendoen-Gort M, *et al*. Systematic review on the measurement properties of diabetes-specific patient-reported outcome measures (PROMs) for measuring physical functioning in people with type 2 diabetes. *BMJ Open Diab Res Care* 2022;**10**:e002729. doi:10.1136/bmjdrc-2021-002729

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjdrc-2021-002729).

Received 17 December 2021
Accepted 9 May 2022

Check for updates

© Author(s) (or their employer(s)) 2022. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to
Dr Caroline B Terwee;
cb.terwee@amsterdamumc.nl

ABSTRACT

We aimed to systematically assess the measurement properties of diabetes-specific patient-reported outcome measures (PROMs) for measuring physical functioning, one of the core outcomes, in adults with type 2 diabetes.

We performed a systematic literature search for PROMs or subscales measuring physical function that were validated to at least some extent in EMBASE and MEDLINE. Measurement properties were evaluated according to the COSMIN guideline for systematic reviews of PROMs.

In total 21 articles were included, describing 12 versions of 7 unique diabetes-specific PROMs or subscales measuring physical functioning. In general, there were few high-quality studies on measurement properties of PROMs measuring physical functioning in adults with type 2 diabetes. The Dependence/Daily Life subscale of the Diabetic Foot Ulcer Scale—Short Form (DFS-SF) and the Impact of Weight on Activities of Daily Living Questionnaire (IWADL) were most extensively evaluated. Both had sufficient ratings for aspects of content validity, although with mostly very low-quality evidence. Sufficient ratings for structural validity, internal consistency, and reliability were also found for both instruments, but responsiveness was rated inconsistent for both instruments. The other PROMs or subscales often had insufficient aspects of content validity, or their unidimensionality could not be confirmed.

This systematic review showed that the Dependence/Daily Life subscale of the DFS-SF and the IWADL could be used to measure physical functioning in people with type 2 diabetes in research or clinical practice, while keeping the limitations of these instruments in mind. The measurement properties that have not been evaluated extensively for these PROMs should be evaluated in future studies.

The study protocol was registered in the PROSPERO database, number CRD42021234890.

INTRODUCTION

The number of adults with diabetes has more than tripled over the past 20 years. In 2019, 463 million adults were estimated to have diabetes, and this number is expected to increase to 700 million in 2045. Around 90% of all diabetes is type 2 diabetes. Ten per cent of global health expenditure is spent on diabetes treatment, making the disease a global problem.[1 2]

Because of the chronic nature of type 2 diabetes and the impact on peoples' lives, it is important to measure patient-reported outcomes, such as symptoms and physical, mental, and social functioning, in research and clinical practice. To that end, patient-reported outcome measures (PROMs) can be used, which measure health outcomes that are important to patients.[3–7] However, a review of type 2 diabetes clinical trials showed that 10% (ie, 14 studies) included patient-reported outcomes.[8] In total, 68 different outcomes were measured in these studies with 23 PROMs. Most PROMs (87%) were used in only one study.[8] This heterogeneity and lack of standardized outcome measurement hampers pooling and comparing outcome data.

To overcome heterogeneity in measured outcomes, a core outcomes set (COS) has been developed for type 2 diabetes.[9] This COS represents an agreed standardized set of outcomes that should be measured and reported in all trials for type 2 diabetes.[9 10] One of the patient-reported outcomes that has been included in the COS for type 2 diabetes is *activities of daily living*, defined as 'being able to complete usual everyday tasks and activities, including those related to personal care, household tasks or community-based tasks.'[9] However, *activities of daily living* does not refer to an aspect of health, as opposed to, for example, *limitations in the performance of activities of daily living*. As such, we personally believe the term *physical functioning*, which often includes activities of daily living,[11] better covers the construct.

BMJ Open Diab Res Care: first published as 10.1136/bmjdrc-2021-002729 on 8 June 2022. Downloaded from http://drc.bmj.com/ on October 15, 2023 by guest. Protected by copyright.

It is important to measure physical functioning with the most suitable PROM, taking specific PROM characteristics into account, such as interpretability of scores (eg, reference values, minimal important change values), feasibility of use, and measurement properties. Measurement properties are the quality aspects of a PROM and include reliability, validity, and responsiveness (see online supplemental appendix 1 for definitions of the measurement properties).[12] To make an evidence-based recommendation on the most suitable PROM, all available PROMs suitable for people with type 2 diabetes need to be evaluated on these characteristics in a high-quality systematic review.

Several systematic reviews on PROMs used in diabetes research have been published in the last decade.[13–22] Most of these reviews included instruments measuring multidimensional constructs,[13 15–19 21 22] but have not reported[13 19] nor evaluated[15–18 21 22] the results per subscale. They also made little to no effort to provide an overview of the different constructs measured by subscales of PROMs. This is important, because the results of measurement properties can vary among subscales and review users need to know what the best instrument is to measure a certain construct. Moreover, several reviews have not conducted a (complete) risk of bias assessment to assess the quality of individual studies[13 17–19] nor have they graded the quality of the total body of evidence for a specific PROM.[13 16–19]

COSMIN (COnsensus-based Standards for the selection of health Measurement INstruments) is the most comprehensive and widespread methodology to enable the evidence-based selection of the most suitable PROM for a certain construct and population.[23] The studies that have used the COSMIN methodology seem to have not correctly applied it, as for example it was unclear how the overall results per PROM were summarized or graded, or this was not done at all.[19 22] Because of these limitations, previous reviews provide limited guidance on which PROMs or subscales are most suitable to measure physical functioning. There is thus still a need for a high-quality systematic review of PROMs for people with diabetes.[24] Therefore, this study aims to systematically assess the measurement properties of diabetes-specific PROMs for measuring physical functioning in adults with type 2 diabetes to make recommendations on the most suitable PROM to use in research or clinical practice.

METHODS

The systematic review was conducted according to the COSMIN guideline for systematic reviews.[23]

Literature search

This study was part of a larger systematic review, in which (1) all PROMs that have been validated to at least some extend in people with type 2 diabetes have been identified and described,[25] (2) the content validity of diabetes-specific PROMs has been investigated,[26] and (3) the measurement properties of diabetes-specific PROMs for physical functioning have been assessed (this study). A comprehensive search was performed in the bibliographic databases MEDLINE (through PubMed) and EMBASE (through www.embase.com) from inception up to January 1, 2022 without language restrictions. Non-English papers were included if relevant information could easily be extracted with Google translate. The search consisted of three elements: (1) type 2 diabetes, using a comprehensive set of search terms from a clinical librarian of the Vrije Universiteit Amsterdam, the Netherlands; (2) PROMs, using a PROM filter[27]; and (3) measurement properties, using a modified version of the measurement properties filter.[28 29] No search terms were used for the construct, as the complete series of reviews intended to find all instruments that have been validated in people with type 2 diabetes. Moreover, for this specific review, we intended to also include physical functioning subscales of PROMs measuring broader constructs, such as quality of life. Adding search terms for physical functioning could have prevented finding these broader instruments as subscales are not always mentioned in the abstract. The complete search strategy can be found in online supplemental appendix 2. Reference lists of included articles were searched by hand to ensure all relevant studies and available translations were considered.[25 26]

Study selection

Covidence[30] was used for screening and selection of abstracts and full-text articles. Relevant articles were selected by first reviewing title and abstract, and if the study seemed relevant or in case of doubt, the full-text article was retrieved and screened. Abstract and full-text screening was done by two reviewers independently. Discrepancies were resolved by discussion and/or consultation of a third reviewer. PROMs that were considered to measure physical functioning based on the Wilson and Cleary model[31] in the first review[25] were included in the current study when the following criteria were met:

1. Construct of interest: The PROM or a relevant subscale of a PROM should measure physical functioning. We adopted the definition of the Patient-Reported Outcomes Measurement Information System (PROMIS), a large US initiative that developed generic PROMs for core health outcomes,[11] which defined physical functioning as the capability to perform physical activities (ie, what a person can do in the daily environment), rather than performance (ie, what a person actually does) or capacity (ie, what a person can do in a standardized-controlled environment, often measured by performance-based tests). Capability to perform physical activities includes the functioning of one's upper extremities (dexterity), lower extremities (walking or mobility), and central regions (neck, back), as well as instrumental activities of daily living, such as running errands. In case a subscale of the

BMJ Open Diab Res Care: first published as 10.1136/bmjdrc-2021-002729 on 8 June 2022. Downloaded from http://drc.bmj.com/ on October 15, 2023 by guest. Protected by copyright.

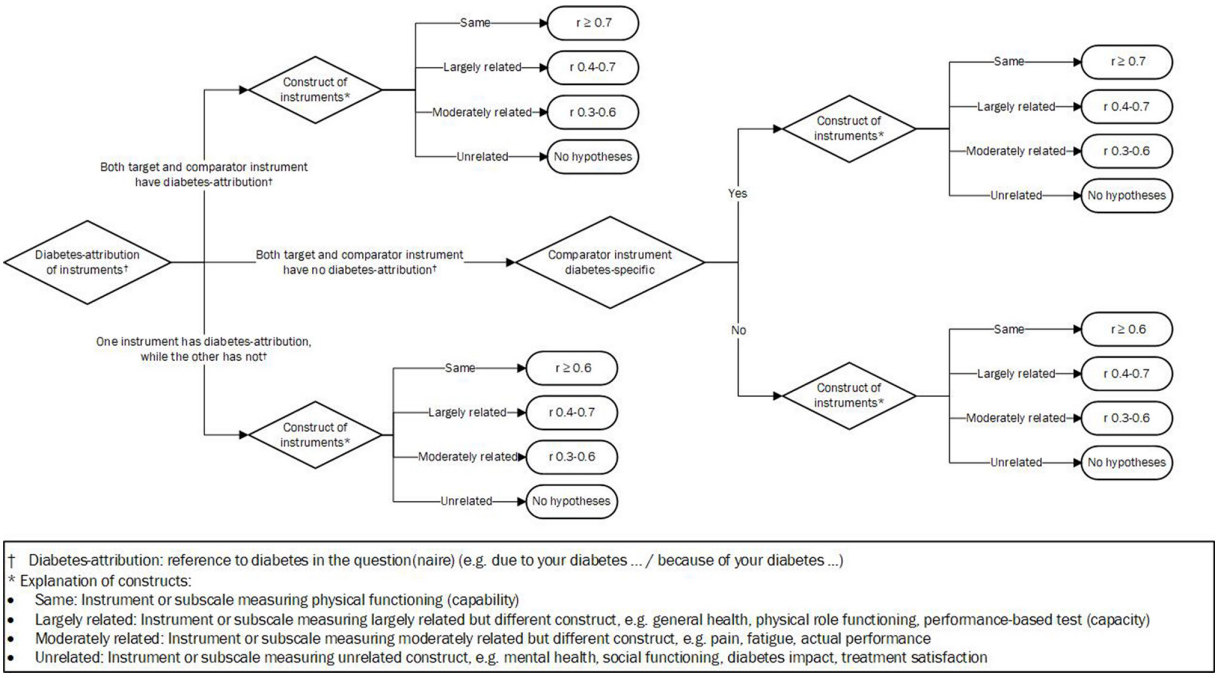


Figure 1 Decision tree for hypotheses regarding the comparisons of instruments.

instrument measures physical functioning, only that subscale was included.

2. Population: At least 50% of the study population or reported subgroups should consist of adults with type 2 diabetes mellitus.
3. Instrument type: The instrument should be a questionnaire, to be completed by the person with type 2 diabetes in self-report or interview form.
4. Measurement properties: At least one of the aims of the paper should be the development of a diabetes-specific PROM or the evaluation of one or more measurement properties of a diabetes-specific PROM. Studies that aim to evaluate the interpretability of a PROM were also included. Studies that use a PROM but do not intend to evaluate its measurement properties or in which the PROM is only used as a comparison instrument in the validation of another instrument were excluded.

Data extraction

PROMs and manuals were retrieved by searching Google or by contacting PROM developers. Characteristics of included PROMs (eg, construct, target population, subscales, number of items, etc.), information on feasibility, and information on interpretability were extracted. For each article, it was determined which measurement properties were evaluated. Data extraction was done by one reviewer and checked by a second reviewer.

Subsequent steps were conducted one measurement property at a time in the following order, as per COSMIN guideline:[23] content validity, internal structure (ie, structural validity, internal consistency, and cross-cultural validity\measurement invariance), reliability and measurement error, and the remaining measurement properties (ie, criterion validity, hypotheses testing for construct validity, and responsiveness). Content validity evidence for the physical functioning subscales was taken from the content validity review,[26] although standard 1, regarding the clarity of the definition of the construct, was scored again specifically for the included physical functioning subscale. Only Dutch or English papers were included for the evaluation of content validity, because this requires detailed understanding of the methods. All other measurement properties were evaluated in the current study.

Evaluation of the quality of a PROM

Per measurement property, first, data on the study population and the results of studies were extracted. Second, the methodological quality of each study was assessed using the COSMIN Risk of Bias checklist.[32] Each standard was rated on a four-point rating scale as 'very good', 'adequate', 'doubtful', or 'inadequate'. A total rating per measurement property per study was obtained taking the lowest rating among the standards (ie, worst-score counts).[33] Third, criteria for good measurement properties were applied to each result using the quality criteria, resulting in a sufficient (+), insufficient (−), or indeterminate (?) rating (online supplemental appendix 3).[23] A priori hypotheses were formulated to evaluate the results on construct validity and responsiveness. Figure 1 shows the predefined hypotheses for comparisons with other instruments. Hypotheses for comparisons between relevant subgroups or before and after intervention were: effect size (eg, Cohen's D, standardized response mean) ≥0.20 for differences between relevant subgroups, score differences between relevant subgroups ≥10% (eg, people with type 2 diabetes should score 10% worse

BMJ Open Diab Res Care: first published as 10.1136/bmjdrc-2021-002729 on 8 June 2022. Downloaded from http://drc.bmj.com/ on October 15, 2023 by guest. Protected by copyright.

than controls), or correlation ≥0.30 between relevant subgroups and score. Relevant subgroups were selected in consultation with an expert on type 2 diabetes. Fourth, evidence from multiple individual studies on the same PROM or subscale was summarized per measurement property and the summarized result was rated against the quality criteria for good measurement properties.[23] The rating of the individual studies (+, −, or ?) was also applied to the summarized result when the results of individual studies were consistent. When individual studies showed inconsistent results, explanations for inconsistency in terms of differences in populations or study quality were explored. When inconsistency could be explained, results were summarized and rated per subset of studies. When inconsistency could not be explained, the overall rating was inconsistent (±), without summarizing the results or based on the majority of consistent results (+, −, or ?). If studies with a + or − rating were available, studies with a ? were ignored and not included when summarizing the results. Fifth, the quality of the evidence was graded using a modified Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach resulting in 'high', 'moderate', 'low', or 'very low' quality.[23] Quality of the evidence was not graded for studies for which the overall rating was indeterminate (?). For all other situations, starting with high-quality evidence, quality of evidence was downgraded (online supplemental appendix 4). For internal consistency, the quality of evidence started at the level of structural validity.[23]

Each step of the quality evaluation was done by two reviewers independently. Discrepancies were resolved by discussion and/or consultation of a third reviewer.

Formulation of recommendations

To formulate recommendations, we considered the results on the measurement properties in order of importance. According to COSMIN, PROMs that have any level of sufficient content validity, which is the most important measurement property, and at least low-quality evidence for sufficient internal consistency (and as such also at least low-quality evidence for sufficient structural validity) can be recommended for use, except when there is high-quality evidence for any insufficient measurement property.[23] We subsequently took results on reliability into account when formulating recommendations, and considered construct validity and responsiveness as least important. Importantly, we also took into account the limitations of the PROMs arising from the recommendations.

RESULTS

Study selection

The database search and reference check resulted in 12 771 unique abstracts, of which 341 were assessed full text for eligibility. Ultimately, 21 articles were included in this review, describing 12 versions of 7 unique PROMs or subscales measuring physical functioning. A flowchart can be found in online supplemental appendix 5. For many PROMs, it was unclear what the PROM exactly aimed to measure, let alone that this was the case for the PROM subscales. In fact, for 7 of the 12 physical functioning subscales, no description was provided at all (table 1). Most PROMs have 5–7 items, although different versions of the included subscale of the Diabetes-39 contain 5–15 items (table 1). Characteristics of study populations involved in PROM design and content validity studies can be found in online supplemental appendix 6, whereas characteristics of study populations for the assessment of other measurement properties can be found in online supplemental appendix 7. Information on feasibility and information on interpretability can be found in online supplemental appendix 8 and online supplemental appendix 9, respectively.

Measurement properties

Table 2 summarizes the results of the included studies on measurement properties per PROM. Per study, the methodological quality and the result of the study are displayed. A more extensive description of the results can be found in online supplemental appendix 10. Table 3 provides an overview of the summary of findings and the quality of the evidence. More extensive results can be found in online supplemental appendix 11. Below, per measurement property the most important results are discussed, in order of importance.[23]

Content validity

The PROM development was considered adequate only for the Diabetic Foot Ulcer Scale—Short Form (DFS-SF),[34] patient-reported outcomes instrument for Thai patients with type 2 diabetes (PRO-DM-Thai),[35] Impact of Weight on Activities of Daily Living Questionnaire (IWADL),[36] and Quality of Life Instrument for Indian Diabetes Patients (QOLID)[37](online supplemental appendix 11). For the other PROMs, the development was rated as inadequate, because the construct of the included physical functioning subscale was not clearly described or the PROM was not pilot tested. Four studies examined the comprehensibility of a PROM (DFS-SF[38] and Diabetes-39[39–41]) after translation (online supplemental appendix 12), which were all doubtful or inadequate. As for most PROMs no or only inadequate content validity studies were available and the PROM development study was also inadequate, the ratings of the relevance, comprehensiveness, and comprehensibility of the PROM were mainly based on the subjective ratings by the reviewers. Note that one reviewer had expertise in PROM development and evaluation (CBT), and one reviewer was a general practitioner and full professor in diabetes care, and as such had expertise in treating people with diabetes (PJME). Considering results of the PROM development studies, content validity studies if both were at least doubtful, and the reviewer ratings, the content validity of the DFS,[42] DFS-SF,[34 38] and IWADL[36]

BMJ Open Diab Res Care: first published as 10.1136/bmjdrc-2021-002729 on 8 June 2022. Downloaded from http://drc.bmj.com/ on October 15, 2023 by guest. Protected by copyright.

Table 1 Characteristics of included diabetes-specific PROMs measuring physical functioning in people with type 2 diabetes (n=11)

PROM—subscale	Construct(s)	Target population	Mode of administration	Recall period	(Sub)scale(s) and number of items*	Original language	Available translations
Diabetic Foot Ulcer Scale (DFS)—Daily activities[42]	PROM: Impact of diabetic foot ulcers on quality of life Included subscale: NR	People with diabetes and foot ulcers	Self-report	NR	11 subscales, 58 items: **Daily activities—six items**	English[42]	Dutch, Danish, Italian, French†
Diabetic Foot Ulcer Scale—Short Form (DFS-SF)—Dependence/daily life[34]	PROM: Impact of diabetic foot ulcers on quality of life Included subscale: Issues related to dependence on others and changes in daily activities	People with diabetes and foot ulcers	Self-report‡	Depending on assessment point, varying from 4 to 20 weeks	Six subscales, 29 items: **Dependence/daily life—five items**	English[34]	Polish,[45] Chinese,[38] Greek,[60] Spanish,[61] Dutch, Danish, Italian, French†
Patient-reported outcomes instrument for Thai patients with type two diabetes (PRO-DM-Thai)—Physical function[35]	PROM: Evaluate outcomes of diabetic care in terms of health and the process of care Included subscale: Relating to physical ability and measuring physical functioning, eg, mobility, dexterity, range of movement, physical activity, activities of daily living	Thai people with type 2 diabetes	Self-report/ interview based	NR	Seven subscales, 44 items: **Physical function—five items**	Thai[35]	
Impact of Weight on Activities of Daily Living Questionnaire (IWADL/APPADL)—(Physical) activities of daily living[36]	PROM (=subscale): Ability to perform daily physical activities	People with type 2 diabetes with moderate obesity (BMI: 30–40)	Self-report	Current	One subscale, seven items: **Physical activities of daily living—seven items**	English[36 46]	
Quality of Life Instrument for Indian Diabetes Patients (QOLID)—Physical endurance[37]	PROM: Quality of life in Indian diabetic people Included subscale: Relate to physical activities	Type 2 diabetes	Interview based	Physical endurance: last 3 months	Eight subscales, 34 items: **Physical endurance—six items**	English/ Hindi§ [37]	
Diabetes Quality of Life Clinical Trial Questionnaire (DQLCTQ)—Physical function[44]	PROM: Quality of life in people with diabetes Included subscale: NR	Type 1 and 2 diabetes	Self-report	NR	Eight subscales, 57 items: **Physical function—six items**	English, German, French[44]	Belgium/Dutch†
Diabetes-39—Energy and Mobility (pilot version, 14 items)[43]	PROM: Quality of life of people with diabetes Included subscale: NR	Diabetes	Self-report	1 month	Six subscales, 42 items: **Energy and Mobility—14 items**	English[43]	
Diabetes-39—Energy and Mobility (15 items)[43]	PROM: Quality of life of people with diabetes Included subscale: NR	Diabetes¶	Self-report‡	1 month	Five subscales, 39 items: **Energy and Mobility—15 items**	English[43]	Arabic,[39] Vietnamese,[40] Portuguese,[41] Spanish,[51] Chinese,[48 59] Danish, Finnish, Norwegian†
MDiabetes-39 Thai—Energy and Mobility (10 items)[49]	PROM: Quality of life of people with diabetes Included subscale: NR	Diabetes	93% interview based, 7% self-report	1 month	Six subscales, 39 items: **Energy and Mobility—10 items**	English,[43] Thai[49]	
Diabetes-39 German—Physical impairment (seven items)[50]	PROM: Quality of life of people with diabetes Included subscale: NR	Type 2 diabetes	Self-report	1 month	Five subscales, 39 items: **Physical impairment—seven items**	English,[43] German[50]	

Continued

BMJ Open Diab Res Care: first published as 10.1136/bmjdrc-2021-002729 on 8 June 2022. Downloaded from http://drc.bmj.com/ on October 15, 2023 by guest. Protected by copyright.

Table 1 Continued

PROM—subscale	Construct(s)	Target population	Mode of administration	Recall period	(Sub)scale(s) and number of items*	Original language	Available translations
Diabetes-39 Short Form—Energy and Mobility (five items)[48]	PROM: Quality of life of people with diabetes Included subscale: NR	Diabetes	Interview based	1 month	Five subscales, 22 items: **Energy and Mobility—five items**	English,[43] Chinese[48 59]	
Chinese Cardiff Wound Impact Schedule (C-CWIS)—Physical symptoms and everyday living[47]	PROM: Disease-specific health-related quality of life in patients with diabetic foot ulcer Included subscale: The impact of symptoms on daily functioning and comfort	People with diabetes and foot ulcers	Self-report and interview based	1 week	Three subscales, 25 items: **Physical symptoms and everyday living—12 items**	Chinese[47]	

*Bold subscales measure physical functioning.
†No publication available.
‡Chinese, Greek, and Spanish version: interview based.
§Language unsure.
¶Vietnamese, Portuguese, Spanish version: type 2 diabetes.
NR, not reported; PROM, patient-reported outcome measure.

for measuring physical functioning was considered sufficient, but often with very low-quality evidence. Details on content validity can be found in our preceding review.[26] Many of the other PROMs had items that were not related to physical functioning. For example at least 8 of the 15 items in the Energy and Mobility subscale of the Diabetes-39 also ask for other health problems, such as loss of vision (n=1), other illnesses (n=3), and energy (n=4).[43] Moreover, the items in the Physical Function subscale of the Diabetes Quality of Life Clinical Trial Questionnaire (DQLCTQ) asks for the duration of limitations, rather than the extent of limitations.[44]

Internal structure

Aspects of internal structure were evaluated for all PROMs or subscales. If studies had inadequate quality for structural validity or cross-cultural validity\measurement invariance, this was often due to small sample sizes. Sufficient structural validity *and* internal consistency was found for the DFS-SF,[34 38 45] PRO-DM-Thai,[35] IWADL,[36 46] and Chinese Cardiff Wound Impact Schedule (C-CWIS).[47] Various factor structures for the subscales of the Diabetes-39 have been found,[39 43 48–50] resulting in different versions with 7, 10, 14, and 15 items of the Energy and Mobility subscale. Only the 14- and 15-item versions were found to be unidimensional with sufficient internal consistency. Internal consistency was considered indeterminate for most other PROMs, despite Cronbach's alpha >0.7, because there was not at least low-quality evidence that the PROMs were unidimensional, which is a prerequisite for correct interpretation of internal consistency. Cross-cultural validity was not evaluated for any of the PROMs, whereas measurement invariance was only assessed for the Dependence/Daily Life subscale of the DFS-SF for the variables sex, age, place of residence, education, type of diabetes, and time since diagnosis.[45] Because only sex impacted one item (depend on others to get out of the house), we rated measurement invariance as sufficient (online supplemental appendix 10).

Reliability and measurement error

Reliability was evaluated for six PROMs or subscales. All studies with inadequate quality had a time interval that was considered to be too long (ie, more than 4 weeks). Sufficient reliability was found for the Dependence subscale of the DFS (but not for the Daily Life subscale),[42] and for the DFS-SF,[34] IWADL,[46] DQLCTQ,[44] and 15-item Energy and Mobility subscale of the Diabetes-39.[40 51] Although reliability was also evaluated for the Physical Impairment subscale of the Diabetes-39, the result could not be rated because it was unclear how the reliability parameter was calculated.[50] Measurement error was evaluated only for the IWADL.[46] The measurement error for using the IWADL in individual persons was rated as insufficient because the smallest detectable change was larger than the minimal important change.

Remaining measurement properties

Figure 2 presents an overview of the evidence on hypotheses testing for construct validity and responsiveness (latter marked with *). Bars reaching within the blue area indicate that the results were in accordance with our predefined hypotheses. Panel A shows the correlations with other instruments based on our decision tree, panel B the percentage score differences between subgroups or before and after intervention, and panel C the effect sizes between subgroups or before and after intervention. Studies with an indeterminate rating are not included in figure 2, because hypotheses were not defined or data were not provided to test the hypotheses. All PROMs have been evaluated with respect to construct validity, except the Energy and Mobility subscale of the Diabetes-39 SF.[48] Most studies were of at least adequate quality. Three studies were of inadequate quality, because they did

BMJ Open Diab Res Care: first published as 10.1136/bmjdrc-2021-002729 on 8 June 2022. Downloaded from http://drc.bmj.com/ on October 15, 2023 by guest. Protected by copyright.

Table 2 Results and quality of studies on measurement properties of diabetes-specific PROMs measuring physical functioning in people with type 2 diabetes

PROM—subscale	Structural validity	Internal consistency	Cross-cultural validity\ measurement invariance	Reliability	Measurement error	Criterion validity	Hypotheses testing for construct validity*	Responsiveness†
DFS—Daily activities[42]	Inadequate ?	Inadequate ?		Inadequate Daily life: – Dependence: +			a. Adequate 4+ / 1– b. Very good ?	c. Doubtful 2–
DFS-SF study 1—Dependence/ daily life[34]	Inadequate ?	Very good +		Inadequate +			a. Adequate 2+ / 3–	c. Very good 1–
DFS-SF study 2—Dependence/ daily life[34]	Very good +	Very good +		Inadequate +			a. Adequate 2+ / 3–	c. Very good 1+
DFS-SF Polish—Dependence/ daily life[45]		Very good +	Inadequate +				a. Very good 4+ / 3–; b1. Inadequate 1+ / 3–; b2. Doubtful ?	
DFS-SF Chinese—Dependence/ daily life[38]		Very good +					a. Very good 5+ / 2– b. Very good 1+ / 1–	
DFS-SF Greek—Dependence/ daily life[60]		Very good +					a. Very good 4+ / 3– b. Very good 14+ / 1–	
DFS-SF Spanish—Dependence/ daily life[61]	Inadequate CFA: – EFA: +	Very good +		Doubtful +			a. Adequate 4+ / 1–	c. Very good 1+
PRO-DM-Thai—Physical function[35]	Very good +	Very good +					b1. Inadequate 1–; b2. Very good ?	
IWADL/APPADL—(Physical) activities of daily living[36]	Doubtful +	Very good +					b. Very good 12+ / 22–	
IWADL/APPADL—(Physical) activities of daily living[46]		Very good +		Adequate +	Doubtful –			d. Very good 2+ / 1–
QOLID—Physical endurance[37]	Inadequate +	Very good ?					a. Adequate 3+ / 3– b. Very good 3–	
DQLCTQ—Physical function[44]		Doubtful ?		Adequate +			b. Very good 1+ / 1–	c. Very good 2–
Diabetes-39—Energy and Mobility (pilot version, 14 items)[43]	Adequate +	Very good +					b. Adequate ?	
Diabetes-39—Energy and Mobility (15 items)[43]	Adequate +	Very good +					a. Adequate 6+ / 4– b. Doubtful ?	
Diabetes-39 Arabic—Energy and Mobility (15 items)[39]	Doubtful +	Very good +					a. Adequate 2+ / 3–	
Diabetes-39 Vietnamese—Energy and mobility (15 items)[40]		Very good +		Doubtful +			b. Very good 4+ / 6–	
Diabetes-39 Portuguese—Energy and Mobility (15 items)[41]		Very good +					b. Very good ?	
Diabetes-39 Spanish—Energy and Mobility (15 items)[51]		Very good +		Doubtful +			b. Very good ?	

Continued

BMJ Open Diab Res Care: first published as 10.1136/bmjdrc-2021-002729 on 8 June 2022. Downloaded from http://drc.bmj.com/ on October 15, 2023 by guest. Protected by copyright.

Table 2 Continued

PROM—subscale	Structural validity	Internal consistency	Cross-cultural validity\ measurement invariance	Reliability	Measurement error	Criterion validity	Hypotheses testing for construct validity*	Responsiveness†
Diabetes-39 Taiwan—Energy and Mobility (15 items)[59]		Very good +					a. Very good ? b. Very good 7+ / 5–	
Diabetes-39 Taiwan—Energy and Mobility (15 items)[48]	Adequate +							
Diabetes-39 Thai—Energy and Mobility (10 items)[49]	Adequate –	Very good ?					a. Adequate 2+ / 3– b. Very good 4+ / 2–	
Diabetes-39 German—Physical impairment (seven items)[50]	Inadequate ?	Doubtful ?		Doubtful ?			a. Very good ? b. Very good 4+	d. Doubtful ?
Diabetes-39 SF—Energy and Mobility (five items)[48]	Very good +							
C-CWIS—Physical symptoms and everyday living[47]	Adequate +	Very good +					a. Inadequate ? b. Inadequate 1+	

+: sufficient rating; –: insufficient rating; ?: indeterminate rating.
*a: comparison with other instruments, b: comparison between subgroups.
†a: comparison to gold standard, b: comparison with other instruments, c: comparison between subgroups, d: before and after intervention.
C-CWIS, Chinese Cardiff Wound Impact Schedule; CFA, confirmatory factor analysis; DFS, Diabetic Foot Ulcer Scale; DFS-SF, Diabetic Foot Ulcer Scale—Short Form; DQLCTQ, Diabetes Quality of Life Clinical Trial Questionnaire; EFA, exploratory factor analysis; IWADL/APPADL, Impact of Weight on Activities of Daily Living Questionnaire; PRO-DM-Thai, patient-reported outcomes instrument for Thai patients with type 2 diabetes; PROMs, patient-reported outcome measures; QOLID, Quality of Life Instrument for Indian Diabetes Patients.

BMJ Open Diab Res Care: first published as 10.1136/bmjdrc-2021-002729 on 8 June 2022. Downloaded from http://drc.bmj.com/ on October 15, 2023 by guest. Protected by copyright.

Table 3 Summary of findings per diabetes-specific PROM measuring physical functioning in people with type 2 diabetes, with a rating of the summarized result (+, –, ±, ?) and a grading of the quality of the evidence (H, M, L, V)

PROM—subscale	Content validity			Structural validity	Internal consistency*	Cross-cultural validity	Measurement invariance	Reliability	Measurement error	Criterion validity	Hypotheses testing for construct validity		Responsiveness			
	Relevance	Comprehensiveness	Comprehensibility								Other instruments	Subgroups	Gold standard	Other instruments	Subgroups	Before–after intervention
DFS—Daily activities[42]	+ (V)	+ (V)	+ (M)	?	?			Daily life: – (V) Dependence: + (V)			+ (M)	?			– (L)	
DFS-SF—Dependence/daily life[34 38 45 60 61]	+ (L)	+ (V)	+ (M)	+ (H)	+ (H)		+ (V)	+ (L)			± (H)	+ (H)			± (H)	
PRO-DM-Thai—Physical function[35]	+ (L)	± (L)	+ (L)	+ (H)	+ (H)							– (V)				
IWADL/APPADL—(Physical) activities of daily living[36 46]	+ (L)	+ (V)	+ (V)	+ (L)	+ (L)			+ (M)	– (L)			± (H)				± (L)
QOLID—Physical endurance[37]	± (V)	?	+ (V)	+ (V)	?						± (V)	– (H)				
DQLCTQ—Physical function[44]	– (V)	?	± (V)		?			+ (L)				± (H)			– (H)	
Diabetes-39—Energy and Mobility (pilot version, 14 items)[43]	– (V)	– (V)	+ (M)	+ (M)	+ (M)							?				
Diabetes-39—Energy and Mobility (15 items)[39–41 43 48 51 59]	– (V)	– (V)	+ (M)	+ (H)	+ (H)			+ (M)			± (H)	± (H)				
Diabetes-39 Thai—Energy and Mobility (10 items)[49]	– (V)	– (V)	+ (M)	– (M)	?						± (M)	± (H)				
Diabetes-39 German—Physical impairment (seven items)[50]	?	?		?	?			?			?	+ (H)				?
Diabetes-39 SF— Energy and Mobility (five items)[48]	– (V)	– (V)	+ (M)	+ (H)												

Continued

BMJ Open Diab Res Care: first published as 10.1136/bmjdrc-2021-002729 on 8 June 2022. Downloaded from http://drc.bmj.com/ on October 15, 2023 by guest. Protected by copyright.

Table 3 Continued

PROM—subscale	Content validity			Structural validity	Internal consistency*	Cross-cultural validity	Measurement invariance	Reliability	Measurement error	Criterion validity	Hypotheses testing for construct validity		Responsiveness			
	Relevance	Comprehensiveness	Comprehensibility								Other instruments	Subgroups	Gold standard	Other instruments	Subgroups	Before–after intervention
C-CWIS—Physical symptoms and everyday living[47]	– (V)	– (V)	+ (L)	+ (M)	+ (M)						?	+ (V)				

+: sufficient overall rating; –: insufficient overall rating; ±: inconsistent overall rating; ?: indeterminate overall rating; V: very low—we have very little confidence in the measurement property estimate, the true measurement property is likely to be substantially different from the estimate of the measurement property; L: low—our confidence in the measurement property is limited, the true measurement property may be substantially different from the estimate of the measurement property; M: moderate—we are moderately confident in the measurement property estimate, the true measurement property is likely to be close to the estimate of the measurement property, but this is a possibility that it is substantially different; H: high—we are very confident that the true measurement property lies close to that of the estimate of the measurement property.
*Per protocol of the COSMIN guideline for systematic reviews, the quality of evidence for internal consistency cannot be higher than the quality of evidence for structural validity.[23]
C-CWIS, Chinese Cardiff Wound Impact Schedule; COSMIN, COnsensus-based Standards for the selection of health Measurement INstruments; DFS, Diabetic Foot Ulcer Scale; DFS-SF, Diabetic Foot Ulcer Scale—Short Form; DQLCTQ, Diabetes Quality of Life Clinical Trial Questionnaire; IWADL/APPADL, Impact of Weight on Activities of Daily Living Questionnaire; PRO-DM-Thai, patient-reported outcomes instrument for Thai patients with type 2 diabetes; PROM, patient-reported outcome measure; QOLID, Quality of Life Instrument for Indian Diabetes Patients.

not apply an appropriate statistical method to compare subgroups.[35 45 47] Construct validity of the Daily Activities subscale of the DFS[42] was considered sufficient based on correlations between instruments, because ≥75% of the results were in accordance with our predefined hypotheses. Construct validity of the Dependence/Daily Life subscale of the DFS-SF, Physical Impairment subscale of the Diabetes-39,[50] and the Physical Symptoms and Everyday living subscale of the C-CWIS were considered sufficient based on comparisons between subgroups, as ≥75% of the results were in accordance with our predefined hypotheses. Responsiveness (marked with *) was evaluated for five PROMs. All studies were of very good quality. For none of the PROMs, responsiveness was considered sufficient.

Recommendations

The DFS-SF and IWADL had sufficient relevance, comprehensiveness, and comprehensibility, and at least low-quality evidence for sufficient internal consistency, and can thus be considered for use in research and clinical practice. Both also had sufficient reliability, but measurement error of the IWADL was insufficient. The DFS-SF and IWADL had inconsistent responsiveness, with high-quality evidence for the subscale of the DFS-SF. This limitation should be taken into account when considering using the DFS-SF and IWADL.

DISCUSSION

This review systematically evaluated the measurement properties of diabetes-specific PROMs for measuring physical functioning, one of the core outcomes,[9] in adults with type 2 diabetes. To ascertain a high-quality systematic review with trustworthy results, we adhered to the COSMIN guideline for systematic reviews.[23] In our review, 12 versions of seven unique PROMs were identified. The Dependence/Daily Life subscale of the DFS-SF[34 38 45] and the IWADL[36 46] seem to be the most extensively evaluated and had sufficient content validity, structural validity, internal consistency, and reliability.

Content validity is the most important measurement property,[23] and the Dependence/Daily Life subscale of the DFS-SF[34 38 45] and the IWADL[36 46] have sufficient relevance, comprehensiveness, and comprehensibility, although mostly based on low or very low quality evidence. The content of the IWADL is more focused on limitations with the performance of daily activities, whereas the content of the Dependence/Daily Life subscale of the DFS-SF asks for dependence on others for the performance of daily activities. Moreover, the DFS-SF was specifically developed for people with type 2 diabetes and foot ulcers. These limitations in content and target population should be taken into account when using the DFS-SF and IWADL. After content validity, structural validity is the second most important measurement property,[23] and both subscales were considered unidimensional. Sufficient internal consistency and reliability were also found

BMJ Open Diab Res Care: first published as 10.1136/bmjdrc-2021-002729 on 8 June 2022. Downloaded from http://drc.bmj.com/ on October 15, 2023 by guest. Protected by copyright.

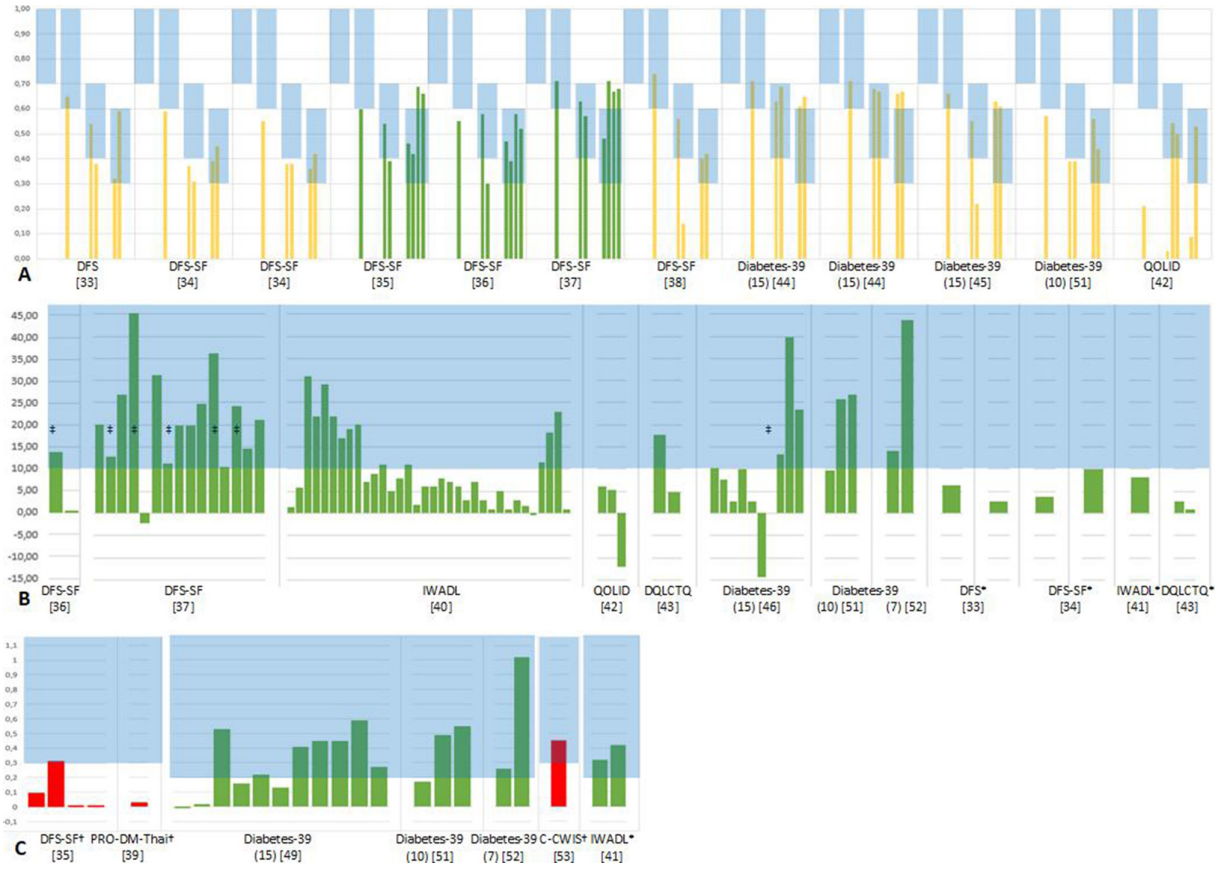


Figure 2 Results on hypotheses testing for construct validity and responsiveness of diabetes-specific PROMs measuring physical functioning in people with type 2 diabetes: (A) correlations with other instruments; (B) Percentage score differences between subgroups or before and after intervention; (C) Effect sizes between subgroups or before and after intervention. *Results of responsiveness; †Correlations between subgroups and instrument score; ‡One of the known-groups tested in the hypotheses was small (n <20); Number in parentheses indicates the number of items in the subscale for the Diabetes-39, for example (15) refers to the 15-item Energy and Mobility subscale; Green: very good study; Yellow: adequate study; Orange: doubtful study; Red: inadequate study; Bars reaching within the blue area indicate that the results are in accordance with our predefined hypotheses, for example, for the DFS in panel A, four results are in accordance with our predefined hypotheses and one is not (one result >0.6, one result 0.4–0.7, two results 0.3–0.6). C-CWIS, Chinese Cardiff Wound Impact Schedule; DFS, Diabetic Foot Ulcer Scale; DFS-SF, Diabetic Foot Ulcer Scale—Short Form; DQLCTQ, Diabetes Quality of Life Clinical Trial Questionnaire; IWADL, Impact of Weight on Activities of Daily Living Questionnaire; PRO-DM-Thai, patient-reported outcomes instrument for Thai patients with type 2 diabetes; PROMs, patient-reported outcome measures; QOLID, Quality of Life Instrument for Indian Diabetes Patients.

for both instruments, although measurement error was insufficient for the IWADL, but the quality of the evidence was low, and therefore further research regarding this measurement property should be conducted. No information about measurement error was available for the Dependence/Daily Life subscale of the DFS-SF. Construct validity in terms of comparisons between subgroups was sufficient for the Dependence/Daily Life subscale of the DFS-SF, whereas this was inconsistent for the IWADL and for correlations between instruments. Responsiveness was also inconsistent for both instruments.

In general, we show in the current review that high-quality studies on measurement properties of PROMs measuring physical functioning in adults with type 2 diabetes are scarce. Five of the studies on PROM development were of inadequate methodological quality, whereas the other four were of doubtful methodological quality. For structural validity, 7 out of 15 studies were of inadequate or doubtful quality and for measurement invariance the one study found was also of inadequate quality. For reliability, six out of eight studies were of inadequate or doubtful quality and for measurement error the one study found was of doubtful quality. The inadequate or doubtful methodological quality of the individual studies resulted in lower quality evidence for many measurement properties. For internal consistency, hypotheses testing for construct validity and responsiveness, the majority of the studies had adequate or very good methodological quality (19 out of 22 for internal consistency, 27 out of 33 for hypotheses testing for construct validity, and 5 out of 7 for responsiveness), leading to higher quality evidence.

Most PROMs or subscales had inconsistent construct validity, often with high-quality evidence, so future studies will probably not change these results. Considering the

BMJ Open Diab Res Care: first published as 10.1136/bmjdrc-2021-002729 on 8 June 2022. Downloaded from http://drc.bmj.com/ on October 15, 2023 by guest. Protected by copyright.

results on comparisons with other instruments, correlations of the DFS-SF Dependence/Daily Life subscale with more related constructs were higher than those with less related construct. Most correlations just not met our hypotheses, which also show that formulating hypotheses is challenging. On the other hand, correlations of the Diabetes-39 15-item Energy and Mobility subscale were all high, regardless of the comparison instrument's construct, indicating that the content of the 15-item Energy and Mobility subscale is not only measuring physical functioning. This also resonates content validity results, with insufficient relevance and comprehensiveness of the Diabetes-39 15-item Energy and Mobility subscale, whereas these were sufficient for the DFS-SF Dependence/Daily Life subscale.

Several PROMs have been translated in various languages, but none of these PROMs have been assessed for cross-cultural validity. This is remarkable, because a large number of PROMs that have originally been developed in English have been translated for use in countries that are likely to be culturally different from western countries, for example, Asian or Arabic countries. Evaluating cross-cultural validity is important, because it is not self-evident that the items in translated questionnaires perform similar compared with the items of the original instrument.[52]

The measurement properties that have not been evaluated for various PROMs could be evaluated in future studies. However, it is not very useful to study these measurement properties for a PROM with insufficient content validity. To measure physical functioning in a valid way, a PROM needs to contain items referring to the functioning of one's upper extremities, lower extremities or central regions, or relevant activities of daily living for people with type 2 diabetes and should not contain items that are not related to physical functioning or that lack key aspects of physical functioning. Only the Dependence/Daily Life subscale of the DFS-SF and the IWADL fulfill these requirements and are worthwhile to be subject of future validation studies.

As an alternative, one could consider using or validating a generic PROM for measuring physical functioning in people with type 2 diabetes. Examples are the Physical Functioning subscale of the 36-item Short Form Health Survey (SF-36), which has been used quite often used in diabetes studies (eg, Refs 53–57), or the more modern, generic PROMIS Physical Function measures.[58] The necessity to use a disease-specific PROM for such a generic outcome as physical functioning can be questioned. It is likely that relevant physical functioning items (eg, walking, stair climbing, performing household activities) are the same for people with diabetes as for people with other conditions. Furthermore, none of the included PROMs had diabetes-attribution in the question(naire) (eg, due to your diabetes…/because of your diabetes…). Diabetes attribution may not always be desirable, because it might lead to differences in interpretation of the items in a PROM. For example, some people will relate the items specifically to their diabetes, while others will ignore the attribution and respond to the items considering their overall health. Also, some people might not know whether their complaints are caused by their diabetes, and as such may doubt how to respond. This may affect reliability and validity of the PROM.

At least eight systematic reviews on PROMs used in diabetes research and care that could potentially have included the same instruments and articles have been published in the last decade.[13 15–19 21 22] However, these reviews included at best only half of the PROMs that were included in this study. Most of these reviews included the Diabetes-39, but often not all language versions were considered. For example, the recent review by Wee *et al*[22] included only the Arabic[39] and Portuguese[41] versions of the Diabetes-39, but not the original English version,[43] nor the Vietnamese,[40] Spanish,[51] Chinese,[48 59] Thai,[49] and German[50] versions. The IWADL[36 46] has not been evaluated in any of the previous reviews. Moreover, most reviews provided a judgment on the quality of only some of the measurement properties. For example, Bottino *et al*[16] only evaluated internal consistency, and content validity and structural validity were almost never evaluated in any of these reviews. Thus, the current review is the first to give a comprehensive overview of the measurement properties of PROMs or subscales that measure physical functioning in people with type 2 diabetes.

A limitation of the current study is that the assessment of content validity was difficult because we included physical functioning subscales from PROMs that often measured broader constructs, such as health-related quality of life. Although the assessment of measurement properties should be conducted for each subscale separately,[23] often information required for the assessment of content validity is only reported for the PROM as a whole. Also for other measurement properties, information was sometimes reported poorly or unclear. Thus, as a team, we had to make decisions on how to value the information. This is inherent to using the COSMIN methodology, but other researchers might come to different conclusions. By reporting everything in tables and appendices, we tried to be as open and consistent as possible.

In conclusion, we identified 12 versions of seven unique diabetes-specific PROMs or subscales for measuring physical functioning, one of the core outcomes in adults with type 2 diabetes. The Dependence/Daily Life subscale of the DFS-SF and the IWADL are most extensively evaluated and had sufficient content validity, structural validity, internal consistency, and reliability, although the quality of the evidence for many measurement properties was very low or low. These PROMs could be used to measure physical functioning in people with type 2 diabetes in research or clinical practice, but the limitations of these instruments (eg, the specific content and target population, inconsistent responsiveness) should be kept in mind. As physical functioning may not necessarily need to be measured with a diabetes-specific PROM, future

BMJ Open Diab Res Care: first published as 10.1136/bmjdrc-2021-002729 on 8 June 2022. Downloaded from http://drc.bmj.com/ on October 15, 2023 by guest. Protected by copyright.

studies should evaluate the validity of generic PROMs, such as PROMIS, in people with type 2 diabetes.

Author affiliations
[1]Department of Epidemiology and Data Science, Amsterdam UMC Locatie VUmc, Amsterdam, The Netherlands
[2]Amsterdam Public Health Research Institute, Amsterdam, Netherlands
[3]Department of General Practice and Elderly Care, Amsterdam UMC Locatie VUmc, Amsterdam, The Netherlands

Acknowledgements We thank Wia Barkema, Lenka Groeneveld, Ilana Halperin, Geetha Mukerji, Amber van der Heijden, and Maartje de Wit for their help with screening titles and abstracts.

Contributors CBT, LBM, EBME, and FR conceived the study; CBT, LBM, and EBME designed the study; CBT carried out the literature search; CBT, PJME, JB, FR, and EBME screened titles and abstracts; CBT, ML-G, EBME, and LBM extracted the data; EBME and LBM assessed the study quality, risk of bias, and grading of the evidence; EBME and LBM interpreted the data; EBME wrote the manuscript; JB, PJME, ML-G, LBM, CBT, and FR revised the manuscript; all authors approved the final version of the manuscript.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data sharing not applicable as no datasets generated and/or analyzed for this study. All data relevant to the study are included in the article or uploaded as supplementary information.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

ORCID iDs
Ellen B M Elsman http://orcid.org/0000-0002-0680-6430
Joline Beulens http://orcid.org/0000-0002-6788-5232

REFERENCES

1 IDF. *Idf diabetes atlas*. 9th Ed. Brussels, Belgium: International Diabetes Federation, 2019. https://www.diabetesatlas.org
2 Unnikrishnan R, Pradeepa R, Joshi SR, *et al*. Type 2 diabetes: Demystifying the global epidemic. *Diabetes* 2017;66:1432–42.
3 Bingham CO, Noonan VK, Auger C, *et al*. Montreal Accord on Patient-Reported Outcomes (PROs) use series - Paper 4: patient-reported outcomes can inform clinical decision making in chronic care. *J Clin Epidemiol* 2017;89:136–41.
4 Noonan VK, Lyddiatt A, Ware P, *et al*. Montreal Accord on Patient-Reported Outcomes (PROs) use series - Paper 3: patient-reported outcomes can facilitate shared decision-making and guide self-management. *J Clin Epidemiol* 2017;89:125–35.
5 Greenhalgh J, Gooding K, Gibbons E, *et al*. How do patient reported outcome measures (PROMs) support clinician-patient communication and patient care? A realist synthesis. *J Patient Rep Outcomes* 2018;2:1–28.
6 Snyder CF, Jensen RE, Segal JB, *et al*. Patient-Reported outcomes (pros): putting the patient perspective in patient-centered outcomes research. *Med Care* 2013;51:S73.
7 Wheat H, Horrell J, Valderas JM, *et al*. Can practitioners use patient reported measures to enhance person centred coordinated care in practice? A qualitative study. *Health Qual Life Outcomes* 2018;16:1–14.
8 Harman NL, James R, Wilding J, *et al*. SCORE-IT (selecting core outcomes for randomised effectiveness trials in type 2 diabetes): a systematic review of registered trials. *Trials* 2017;18:1–13.
9 Harman NL, Wilding JPH, Curry D, *et al*. Selecting core outcomes for randomised effectiveness trials in type 2 diabetes (SCORE-IT): a patient and healthcare professional consensus on a core outcome set for type 2 diabetes. *BMJ Open Diabetes Res Care* 2019;7:e000700.
10 Williamson PR, Altman DG, Bagley H, *et al*. The comet Handbook: version 1.0. *Trials* 2017;18:1–50.
11 HealthMeasures. Promis physical function scoring manual 2020.
12 Mokkink LB, Terwee CB, Patrick DL, *et al*. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol* 2010;63:737–45.
13 Roborel de Climens A, Tunceli K, Arnould B, *et al*. Review of patient-reported outcome instruments measuring health-related quality of life and satisfaction in patients with type 2 diabetes treated with oral therapy. *Curr Med Res Opin* 2015;31:643–65.
14 van Dijk SEM, Adriaanse MC, van der Zwaan L, *et al*. Measurement properties of depression questionnaires in patients with diabetes: a systematic review. *Qual Life Res* 2018;27:1415–30.
15 Vieta A, Badia X, Sacristán JA. A systematic review of patient-reported and economic outcomes: value to stakeholders in the decision-making process in patients with type 2 diabetes mellitus. *Clin Ther* 2011;33:1225–45.
16 Bottino LG, Madalosso MM, Garcia SP, *et al*. Diabetes-Specific questionnaires validated in Brazilian Portuguese: a systematic review. *Arch Endocrinol Metab* 2020;64:111–20.
17 Hogg FRA, Peach G, Price P, *et al*. Measures of health-related quality of life in diabetes-related foot disease: a systematic review. *Diabetologia* 2012;55:552–65.
18 Levterova BA, Dimitrova DD, Levterov GE, *et al*. Instruments for disease-specific quality-of-life measurement in patients with type 2 diabetes mellitus--a systematic review. *Folia Med* 2013;55:83–92.
19 Ortega-Avila AB, Cervera-Garvi P, Ramos-Petersen L, *et al*. Patient-Reported outcome measures for patients with diabetes mellitus associated with foot and ankle pathologies: a systematic review. *J Clin Med* 2019;8. doi:10.3390/jcm8020146. [Epub ahead of print: 27 Jan 2019].
20 Lee J, Lee E-H, Kim C-J, *et al*. Diabetes-Related emotional distress instruments: a systematic review of measurement properties. *Int J Nurs Stud* 2015;52:1868–78.
21 Chen YT, Tan YZ, Cheen M, *et al*. Patient-Reported outcome measures in registry-based studies of type 2 diabetes mellitus: a systematic review. *Curr Diab Rep* 2019;19:135.
22 Wee PJL, Kwan YH, Loh DHF, *et al*. Measurement properties of patient-reported outcome measures for diabetes: systematic review. *J Med Internet Res* 2021;23:e25002.
23 Prinsen CAC, Mokkink LB, Bouter LM, *et al*. COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual Life Res* 2018;27:1147–57.
24 Rutters F, Elsman E, Groeneveld L, *et al*. Challenges in Measuring What Matters to Patients With Diabetes. Comment on "Measurement Properties of Patient-Reported Outcome Measures for Diabetes: Systematic Review". *J Med Internet Res* 2022;24:e36876.
25 Langendoen-Gort M, Groeneveld L, Prinsen CA. Patient-Reported outcome measures for assessing health-related quality of life in people with type 2 diabetes: a systematic review 2021.
26 Terwee CB, Elders P, Langendoen-Gort M. Content validity of patient-reported outcome measures developed for assessing health-related quality of life in people with type 2 diabetes mellitus. *Curre Diab Rep* 2022.
27 Mackintosh A, Hadi M. Prom Group Construct & Instrument Type Filers Oxford, UK: Patient-reported outcome measurement group, 2010. Available: https://cosmin.nl/wp-content/uploads/prom-search-filter-oxford-2010.pdf
28 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;18:1115–23.
29 COSMIN. Search filters, 2021. Available: https://www.cosmin.nl/tools/pubmed-search-filters/
30 Covidence. Covidence: better systematic review management, 2021. Available: https://www.covidence.org/

BMJ Open Diab Res Care: first published as 10.1136/bmjdrc-2021-002729 on 8 June 2022. Downloaded from http://drc.bmj.com/ on October 15, 2023 by guest. Protected by copyright.

31 Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. *JAMA* 1995;273:59–65.
32 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;27:1171–9.
33 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;21:651–7.
34 Bann CM, Fehnel SE, Gagnon DD. Development and validation of the diabetic foot ulcer Scale-short form (DFS-SF). *Pharmacoeconomics* 2003;21:1277–90.
35 Chuayruang K, Sriratanaban J, Hiransuthikul N, *et al*. Development of an instrument for patient-reported outcomes in Thai patients with type 2 diabetes mellitus (PRO-DM-Thai). *Asian Biomedicine* 2017;9:7–19.
36 Hayes RP, Nelson DR, Meldahl ML, *et al*. Ability to perform daily physical activities in individuals with type 2 diabetes and moderate obesity: a preliminary validation of the impact of weight on activities of daily living questionnaire. *Diabetes Technol Ther* 2011;13:705–12.
37 Nagpal J, Kumar A, Kakar S, *et al*. The development of 'Quality of Life Instrument for Indian Diabetes patients (QOLID): a validation and reliability study in middle and higher income groups. *J Assoc Physicians India* 2010;58:295–304.
38 Hui LF, Yee-Tak Fong D, Yam M, *et al*. Translation and validation of the chinese diabetic foot ulcer scale - short form. *Patient* 2008;1:137–45.
39 Khader YS, Bataineh S, Batayha W. The Arabic version of Diabetes-39: psychometric properties and validation. *Chronic Illn* 2008;4:257–63.
40 Nguyen TQ, Vo TQ, Nguyen GH, *et al*. Assessment of health-related quality of life in patients with type II diabetes mellitus: a population-based study at a tertiary hospital. *JCDR* 2018;12:LC44–51.
41 Queiroz FAde, Pace AE, Santos CBdos. Cross-cultural adaptation and validation of the instrument Diabetes - 39 (D-39): brazilian version for type 2 diabetes mellitus patients - stage 1. *Rev Lat Am Enfermagem* 2009;17:708–15.
42 Abetz L, Sutton M, Brady L, *et al*. The diabetic foot ulcer scale (DFS): a quality of life instrument for use in clinical trials. *Prac Diabe Intern* 2002;19:167–75.
43 Boyer JG, Earp JA. The development of an instrument for assessing the quality of life of people with diabetes. Diabetes-39. *Med Care* 1997;35:440–53.
44 Shen W, Kotsanos JG, Huster WJ, *et al*. Development and validation of the diabetes quality of life clinical trial questionnaire. *Med Care* 1999;37:AS45–66.
45 Macioch T, Sobol E, Krakowiecki A, *et al*. Health related quality of life in patients with diabetic foot ulceration - translation and Polish adaptation of Diabetic Foot Ulcer Scale short form. *Health Qual Life Outcomes* 2017;15:15.
46 Hayes RP, Schultz EM, Naegeli AN, *et al*. Test-Retest, responsiveness, and minimal important change of the ability to perform physical activities of daily living questionnaire in individuals with type 2 diabetes and obesity. *Diabetes Technol Ther* 2012;14:1118–25.
47 Huang Y, Wu M, Xing P, *et al*. Translation and validation of the Chinese Cardiff wound impact schedule. *Int J Low Extrem Wounds* 2014;13:5–11.
48 Leite WL, Huang I-C, Marcoulides GA. Item selection for the development of short forms of scales using an ant colony optimization algorithm. *Multivariate Behav Res* 2008;43:411–31.
49 Songraksa K, Lerkiatbundit S. Development of a disease specific quality of life instrument: Thai version of the Diabetic-39. *Song Medic Jou* 2009;27:35–49.
50 Hirsch A, Bartholomae C, Volmer T. Dimensions of quality of life in people with non-insulin-dependent diabetes. *Qual Life Res* 2000;9:207–18.
51 López-Carmona JM, Rodríguez-Moctezuma R. [Adaptation and validation of quality of life instrument Diabetes 39 for Mexican patients with type 2 diabetes mellitus]. *Salud Publica Mex* 2006;48:200–11.
52 Krogsgaard MR, Brodersen J, Christensen KB, *et al*. How to translate and locally adapt a PROM. assessment of cross-cultural differential item functioning. *Scand J Med Sci Sports* 2021;31:999–1008.
53 Anderson RM, Fitzgerald JT, Wisdom K, *et al*. A comparison of global versus disease-specific quality-of-life measures in patients with NIDDM. *Diabetes Care* 1997;20:299–305.
54 Glasziou P, Alexander J, Beller E, *et al*. Which health-related quality of life score? A comparison of alternative utility measures in patients with type 2 diabetes in the advance trial. *Health Qual Life Outcomes* 2007;5:21.
55 Linzer M, Pierce C, Lincoln E, *et al*. Preliminary validation of a patient-based self-assessment measure of severity of illness in type 2 diabetes: results from the pilot phase of the Veterans health study. *J Ambul Care Manage* 2005;28:167–76.
56 Woodcock AJ, Julious SA, Kinmonth AL, *et al*. Problems with the performance of the SF-36 among people with type 2 diabetes in general practice. *Qual Life Res* 2001;10:661–70.
57 Yordanova S, Petkova V, Petrova G, *et al*. Comparison of health-related quality-of-life measurement instruments in diabetic patients. *Biotechnol Biotechnol Equip* 2014;28:769–74.
58 Cella D, Riley W, Stone A, *et al*. The patient-reported outcomes measurement information system (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. *J Clin Epidemiol* 2010;63:1179–94.
59 Huang I-C, Hwang C-C, Wu M-Y, *et al*. Diabetes-specific or generic measures for health-related quality of life? Evidence from psychometric validation of the D-39 and SF-36. *Value Health* 2008;11:450–61.
60 Kontodimopoulos N, Veniou A, Tentolouris N, *et al*. Validity and reliability of the Greek version of the Diabetic Foot Ulcer Scale - Short Form (DFS-SF). *Hormones* 2016;15:394–403.
61 Martinez-Gonzalez D, Dòria M, Martínez-Alonso M, *et al*. Adaptation and validation of the diabetic foot ulcer Scale-Short form in Spanish subjects. *J Clin Med* 2020;9:2497.

BMJ Open Diab Res Care: first published as 10.1136/bmjdrc-2021-002729 on 8 June 2022. Downloaded from http://drc.bmj.com/ on October 15, 2023 by guest. Protected by copyright.

Appendix 1. COSMIN definitions of domains, measurement properties and aspects of measurement properties[12]

Domain	Term: Measurement property	Term: Measurement property aspect	Definition
Reliability			The degree to which the measurement is free from measurement error
Reliability (extended definition)			The extent to which scores for patients who have not changed are the same for repeated measurement under several conditions: e.g., using different sets of items from the same OMI (internal consistency); over time (test-retest); by different persons on the same occasion (inter-rater); or by the same persons on different occasions (intra-rater)
	Internal consistency		The degree of interrelatedness among the items
	Reliability		The proportion of the total variance in the measurements which is due to 'true'† differences between patients
	Measurement error		The systematic and random error of a patient's score that is not attributed to true changes in the construct to be measured
Validity			The degree to which an OMI measures the construct(s) it purports to measure
	Content validity		The degree to which the content of an OMI is an adequate reflection of the construct to be measured
		Face validity	The degree to which (the items of) an OMI indeed seems to be an adequate reflection of the construct to be measured
	Construct validity		The degree to which the scores of an OMI are consistent with hypotheses (e.g., with regard to internal relationships, relationships to scores of other OMIs, or differences between relevant groups) based on the assumption that the OMI validly measures the construct to be measured
		Structural validity	The degree to which the scores of an OMI are an adequate reflection of the dimensionality of the construct to be measured
		Hypotheses testing	Idem construct validity
		Cross-cultural validity	The degree to which the performance of the items on a translated or culturally adapted OMI are an adequate reflection of the performance of the items of the original version of the OMI
	Criterion validity		The degree to which the scores of an OMI are an adequate reflection of a gold standard
Responsiveness			The ability of an OMI to detect change over time in the construct to be measured
	Responsiveness		Idem responsiveness
Interpretability*			The degree to which one can assign qualitative meaning (i.e., clinical or commonly understood connotations) to an OMI's quantitative scores or change in scores

COSMIN: COnsensus-based Standards for the selection of health Measurement INstruments

* Not considered a measurement property, but an important characteristic of a measurement instrument

† The word "true" must be seen in the context of the CTT, which states that any observation is composed of two components—a true score and error associated with the observation. "True" is the average score that would be obtained if the scale were given an infinite number of times. It refers only to the consistency of the score and not to its accuracy.[70]

Appendix 2. Search strategy

PUBMED search January 1, 2022

#1 Diabetes type 2

((Diabet*[tiab] AND (("non insulin"[tiab] AND depend*[tiab]) OR ("noninsulin"[tiab] AND depend*[tiab]) OR "type 2"[tiab] OR "type II" [tiab])) OR iddm[tiab] OR niddm[tiab] OR "glucose intolerance"[tiab] OR "insulin resistant"[tiab] OR "insulin resistance"[tiab])

#2 Modified filter for studies on measurement properties*

~~instrumentation[sh] OR methods[sh] OR~~ "Validation Studies"[pt] ~~OR "Comparative Study"[pt]~~ OR "psychometrics"[MeSH] OR psychometr*[tiab] OR clinimetr*[tw] OR clinometr*[tw] OR "outcome assessment (health care)"[MeSH] OR "outcome assessment"[tiab] OR "outcome measure*"[tw] OR "observer variation"[MeSH] OR "observer variation"[tiab] ~~OR "Health Status Indicators"[Mesh]~~ OR "reproducibility of results"[MeSH] OR reproducib*[tiab] OR "discriminant analysis"[MeSH] OR reliab*[tiab] OR unreliab*[tiab] OR valid*[tiab] OR "coefficient of variation"[tiab] ~~OR coefficient[tiab]~~ OR homogeneity[tiab] OR homogeneous[tiab] OR "internal consistency"[tiab] OR (cronbach*[tiab] AND (alpha[tiab] OR alphas[tiab])) OR (item[tiab] AND (correlation*[tiab] OR selection*[tiab] OR reduction*[tiab])) OR agreement[tw] OR precision[tw] OR imprecision[tw] OR "precise values"[tw] OR test-retest[tiab] OR (test[tiab] AND retest[tiab]) OR (reliab*[tiab] AND (test[tiab] OR retest[tiab])) OR stability[tiab] OR interrater[tiab] OR inter-rater[tiab] OR intrarater[tiab] OR intra-rater[tiab] OR intertester[tiab] OR inter-tester[tiab] OR intratester[tiab] OR intra-tester[tiab] OR interobserver[tiab] OR inter-observer[tiab] OR intraobserver[tiab] OR intra-observer[tiab] OR intertechnician[tiab] OR inter-technician[tiab] OR intratechnician[tiab] OR intra-technician[tiab] OR interexaminer[tiab] OR inter-examiner[tiab] OR intraexaminer[tiab] OR intra-examiner[tiab] OR interassay[tiab] OR inter-assay[tiab] OR intraassay[tiab] OR intra-assay[tiab] OR interindividual[tiab] OR inter-individual[tiab] OR intraindividual[tiab] OR intra-individual[tiab] OR interparticipant[tiab] OR inter-participant[tiab] OR intraparticipant[tiab] OR intra-participant[tiab] OR kappa[tiab] OR kappa's[tiab] OR kappas[tiab] OR repeatab*[tw] OR ((replicab*[tw] OR repeated[tw]) AND (measure[tw] OR measures[tw] OR findings[tw] ~~OR result[tw] OR results[tw]~~ OR test[tw] OR tests[tw])) OR generaliza*[tiab] OR generalisa*[tiab] OR concordance[tiab] OR (intraclass[tiab] AND correlation*[tiab]) OR discriminative[tiab] OR "known group"[tiab] OR "factor analysis"[tiab] OR "factor analyses"[tiab] OR "factor structure"[tiab] OR "factor structures"[tiab] ~~OR dimension*[tiab]~~ OR subscale*[tiab] OR (multitrait[tiab] AND scaling[tiab] AND (analysis[tiab] OR analyses[tiab])) OR "item discriminant"[tiab] OR "interscale correlation*"[tiab] OR error[tiab] OR errors[tiab] OR "individual variability"[tiab] OR "interval variability"[tiab] OR "rate variability"[tiab] OR (variability[tiab] AND (analysis[tiab] OR values[tiab])) ~~OR (uncertainty[tiab] AND (measurement[tiab] OR measuring[tiab]))~~ OR "standard error of measurement"[tiab] ~~OR sensitiv*[tiab]~~ OR responsive*[tiab] OR (limit[tiab] AND detection[tiab]) OR "minimal detectable concentration"[tiab] OR interpretab*[tiab] OR ((minimal[tiab] OR minimally[tiab] OR clinical[tiab] OR clinically[tiab]) AND (important[tiab] ~~OR significant[tiab]~~ OR detectable[tiab]) AND (change[tiab] OR difference[tiab])) OR (small*[tiab] AND (real[tiab] OR detectable[tiab]) AND (change[tiab] OR difference[tiab])) OR "meaningful change"[tiab] OR "ceiling effect"[tiab] OR "floor effect"[tiab] OR "Item response model"[tiab] OR IRT[tiab] OR Rasch[tiab] OR "Differential item functioning"[tiab] OR DIF[tiab] OR "computer adaptive testing"[tiab] OR "item bank"[tiab] OR "cross-cultural equivalence"[tiab]

#3 PROM filter (developed by the University of Oxford, see www.comin.nl)

(HR-PRO[tiab] OR HRPRO[tiab] OR HRQL[tiab] OR HRQoL[tiab] OR QL[tiab] OR QoL[tiab] OR quality of life[tw] OR life quality[tw] OR health index*[tiab] OR health indices[tiab] OR health profile*[tiab] OR health status[tw] OR ((patient[tiab] OR self[tiab] OR child[tiab] OR parent[tiab] OR carer[tiab] OR proxy[tiab]) AND ((report[tiab] OR reported[tiab] OR reporting[tiab]) OR (rated[tiab] OR rating[tiab] OR ratings[tiab]) OR based[tiab] OR (assessed[tiab] OR assessment[tiab] OR assessments[tiab]))) OR ((disability[tiab] OR function[tiab] OR functional[tiab] OR functions[tiab] OR subjective[tiab] OR utility[tiab] OR utilities[tiab] OR wellbeing[tiab] OR well being[tiab]) AND (index[tiab] OR indices[tiab] OR instrument[tiab] OR instruments[tiab] OR measure[tiab] OR measures[tiab] OR questionnaire[tiab] OR questionnaires[tiab] OR profile[tiab] OR profiles[tiab] OR scale[tiab] OR scales[tiab] OR score[tiab] OR scores[tiab] OR status[tiab] OR survey[tiab] OR surveys[tiab])))

(#1 AND #2 AND #3) NOT ("addresses"[Publication Type] OR "biography"[Publication Type] OR "case reports"[Publication Type] OR "comment"[Publication Type] OR "directory"[Publication Type] OR "editorial"[Publication Type] OR "festschrift"[Publication Type] OR "interview"[Publication Type] OR "lectures"[Publication Type] OR "legal cases"[Publication Type] OR "legislation"[Publication Type] OR "letter"[Publication Type] OR "news"[Publication Type] OR "newspaper article"[Publication Type] OR "patient education handout"[Publication Type] OR "popular works"[Publication Type] OR "congresses"[Publication Type] OR "consensus development conference"[Publication Type] OR "consensus development conference, nih"[Publication Type] OR "practice guideline"[Publication Type]) NOT ("animals"[MeSH Terms] NOT "humans"[MeSH Terms])

EMBASE search January 1, 2022

#1 diabetes type 2

(Diabet*:ti,ab AND (('non insulin':ti,ab AND depend*:ti,ab) OR (noninsulin:ti,ab AND depend*:ti,ab) OR 'type 2':ti,ab OR 'type II':ti,ab)) OR iddm:ti,ab OR niddm:ti,ab OR 'glucose intolerance':ti,ab OR 'insulin resistant':ti,ab OR 'insulin resistance':ti,ab

#2 Modified filter for studies on measurement properties*

~~'intermethod comparison'/exp OR~~ 'data collection method'/exp OR 'validation study'/exp OR 'feasibility study'/exp OR 'pilot study'/exp OR 'psychometry'/exp OR 'reproducibility'/exp OR reproducib*:ab,ti OR 'audit':ab,ti OR psychometr*:ab,ti OR clinimetr*:ab,ti OR clinometr*:ab,ti OR 'observer variation'/exp OR 'observer variation':ab,ti OR 'discriminant analysis'/exp OR 'validity'/exp OR reliab*:ab,ti OR valid*:ab,ti ~~OR 'coefficient':ab,ti~~ OR 'internal consistency':ab,ti OR (cronbach*:ab,ti AND ('alpha':ab,ti OR 'alphas':ab,ti)) OR 'item correlation':ab,ti OR 'item correlations':ab,ti OR 'item selection':ab,ti OR 'item selections':ab,ti OR 'item reduction':ab,ti OR 'item reductions':ab,ti OR 'agreement':ab,ti OR 'precision':ab,ti OR 'imprecision':ab,ti OR 'precise values':ab,ti OR 'test-retest':ab,ti OR ('test':ab,ti AND 'retest':ab,ti) OR (reliab*:ab,ti AND ('test':ab,ti OR 'retest':ab,ti)) OR 'stability':ab,ti OR 'interrater':ab,ti OR 'inter-rater':ab,ti OR 'intrarater':ab,ti OR 'intra-rater':ab,ti OR 'intertester':ab,ti OR 'inter-tester':ab,ti OR 'intratester':ab,ti OR 'intra-tester':ab,ti OR 'interobeserver':ab,ti OR 'inter-observer':ab,ti OR 'intraobserver':ab,ti OR 'intra-observer':ab,ti OR 'intertechnician':ab,ti OR 'inter-technician':ab,ti OR 'intratechnician':ab,ti OR 'intra-technician':ab,ti OR 'interexaminer':ab,ti OR 'inter-examiner':ab,ti OR 'intraexaminer':ab,ti OR 'intra-examiner':ab,ti OR 'interassay':ab,ti OR 'inter-assay':ab,ti OR 'intraassay':ab,ti OR 'intra-assay':ab,ti OR 'interindividual':ab,ti OR 'inter-individual':ab,ti OR 'intraindividual':ab,ti OR 'intra-individual':ab,ti OR 'interparticipant':ab,ti OR 'inter-participant':ab,ti OR 'intraparticipant':ab,ti OR 'intra-participant':ab,ti OR 'kappa':ab,ti OR 'kappas':ab,ti OR 'coefficient of variation':ab,ti OR repeatab*:ab,ti OR (replicab*:ab,ti OR 'repeated':ab,ti AND ('measure':ab,ti OR 'measures':ab,ti OR 'findings':ab,ti ~~OR 'result':ab,ti OR 'results':ab,ti~~ OR 'test':ab,ti OR 'tests':ab,ti)) OR generaliza*:ab,ti OR generalisa*:ab,ti OR 'concordance':ab,ti OR ('intraclass':ab,ti AND correlation*:ab,ti) OR 'discriminative':ab,ti OR 'known group':ab,ti OR 'factor analysis':ab,ti OR 'factor analyses':ab,ti OR 'factor structure':ab,ti OR 'factor structures':ab,ti OR 'dimensionality':ab,ti OR subscale*:ab,ti OR 'multitrait scaling analysis':ab,ti OR 'multitrait scaling analyses':ab,ti OR 'item discriminant':ab,ti OR 'interscale correlation':ab,ti OR 'interscale correlations':ab,ti OR ('error':ab,ti OR 'errors':ab,ti AND (measure*:ab,ti OR correlat*:ab,ti OR evaluat*:ab,ti OR 'accuracy':ab,ti OR 'accurate':ab,ti OR 'precision':ab,ti OR 'mean':ab,ti)) OR 'individual variability':ab,ti OR 'interval variability':ab,ti OR 'rate variability':ab,ti OR 'variability analysis':ab,ti OR 'standard error of measurement':ab,ti ~~OR sensitiv*:ab,ti~~ OR responsive*:ab,ti OR ('limit':ab,ti AND 'detection':ab,ti) OR 'minimal detectable concentration':ab,ti OR interpretab*:ab,ti OR (small*:ab,ti AND ('real':ab,ti OR 'detectable':ab,ti) AND ('change':ab,ti OR 'difference':ab,ti)) OR 'meaningful change':ab,ti OR 'minimal important change':ab,ti OR 'minimal important difference':ab,ti OR 'minimally important change':ab,ti OR 'minimally important difference':ab,ti OR 'minimal detectable change':ab,ti OR 'minimal detectable difference':ab,ti OR 'minimally detectable change':ab,ti OR 'minimally detectable difference':ab,ti OR 'minimal real change':ab,ti OR 'minimal real difference':ab,ti OR 'minimally real change':ab,ti OR 'minimally real difference':ab,ti OR 'ceiling effect':ab,ti OR 'floor effect':ab,ti OR 'item response model':ab,ti OR 'irt':ab,ti OR 'rasch':ab,ti OR 'differential item functioning':ab,ti OR 'dif':ab,ti OR 'computer adaptive testing':ab,ti OR 'item bank':ab,ti OR 'cross-cultural equivalence':ab,ti

#3 PROM filter (developed by the University of Oxford, see www.comin.nl)

(HR-PRO:ti,ab OR HRPRO:ti,ab OR HRQL:ti,ab OR HRQoL:ti,ab OR QL:ti,ab OR QoL:ti,ab OR 'quality of life':ti,ab OR 'life quality':ti,ab OR 'health index*':ti,ab OR 'health indices':ti,ab OR 'health profile*':ti,ab OR 'health status':ti,ab OR ((patient:ti,ab OR self:ti,ab OR child:ti,ab OR parent:ti,ab OR carer:ti,ab OR proxy:ti,ab) AND ((report:ti,ab OR reported:ti,ab OR reporting:ti,ab) OR (rated:ti,ab OR rating:ti,ab OR ratings:ti,ab) OR based:ti,ab OR (assessed:ti,ab OR assessment:ti,ab OR assessments:ti,ab))) OR ((disability:ti,ab OR function:ti,ab OR functional:ti,ab OR functions:ti,ab OR subjective:ti,ab OR utility:ti,ab OR utilities:ti,ab OR wellbeing:ti,ab OR 'well being':ti,ab) AND (index:ti,ab OR indices:ti,ab OR instrument:ti,ab OR instruments:ti,ab OR measure:ti,ab OR measures:ti,ab OR questionnaire:ti,ab OR questionnaires:ti,ab OR profile:ti,ab OR profiles:ti,ab OR scale:ti,ab OR scales:ti,ab OR score:ti,ab OR scores:ti,ab OR status:ti,ab OR survey:ti,ab OR surveys:ti,ab)))

#4 publicatie types

#3 AND ('article'/it OR 'article in press'/it OR 'review'/it)

#5 not animals

#4 NOT ([animals]/lim NOT [humans]/lim)

* Modified from Terwee et al. [28]. The crossed out search terms were left out because these terms, in combination with the search terms for diabetes, yielded too many abstracts to read.

Appendix 3. Criteria for good measurement properties

Measurement property	Rating	Criteria
Structural validity*	+	**CTT:** EFA/PCA: factor loadings of each item on its factor is at least 0.30 *AND* maximum 10% of the items load on more than one factor *AND* minimum explained variance is 50% and structure is in line with the theory about the construct to be measured *OR* results on scree plot or Kaiser criterion (Eigenvalues >1) are in line with the theory about the construct to be measured CFA: CFI or TLI or comparable measure >0.95 *OR* RMSEA <0.06 *OR* SRMR <0.08 **IRT/Rasch:** no violation of unidimensionality: CFI or TLI or comparable measure >0.95 *OR* RMSEA <0.06 *OR* SRMR <0.08 *AND* no violation of local independence: residual correlations among the items after controlling for dominant factor <0.20 *OR* Q3's <0.37 *AND* no violation of monotonicity: adequate looking graphs *OR* item scalability >0.30 *AND* adequate model fit: IRT: χ^2 >0.01 Rasch: infit and outfit mean squares ≥0.5 and ≤1.5 *OR* Z-standardized values >-2 and <2
	?	**CTT:** not all information for '+' reported **IRT/Rasch:** model fit not reported
	-	Criteria for '+' not met
Internal consistency	+	At least low evidence for sufficient structural validity *AND* Cronbach's alpha(s) ≥0.70 for each unidimensional scale or subscale
	?	Criteria for "at least low eveidence for sufficient structural validity" not met
	-	At leats low evidence for sufficient structural validity *AND* Cronbach's alpha(s) <0.70 for each unidimensional scale or subscale
Reliability	+	ICC or (weighthed) kappa or Pearson/Spearman correlation ≥0.70
	?	ICC or (weighthed) kappa or Pearson/Spearman correlation not reported
	-	ICC or (weighthed) kappa or Pearson/Spearman correlation <0.70
Measurement error	+	SDC or LoA <MIC
	?	MIC not defined
	-	SDC or LoA > MIC
Hypotheses testing for construct validity	+	≥75% of the results is in accordance with predefined hypotheses
	?	No hypotheses defined (by the review team)
	-	≥75% of the results is not in accordance with predefined hypotheses
Cross-cultural validity\ measurement invariance	+	No important differences found between group factors (such as age, gender, language) in multiple group factor analysis *OR* no important DIF for group factors (McFadden's R^2 <0.02)
	?	No multiple group factor analysis *OR* DIF analysis performed
	-	Important differences between group factors *OR* DIF was found
Criterion validity	+	Correlation with gold standard ≥0.70 *OR* AUC ≥0.70
	?	Not all information for '+' reported
	-	Correlation with gold standard <0.70 *OR* AUC <0.70
Responsiveness	+	≥75% of the results is in accordance with predefined hypotheses *OR* AUC ≥0.70

	?	No hypotheses defined (by the review team)
	-	≥75% of the results is not in accordance with predefined hypotheses *OR* AUC <0.70

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

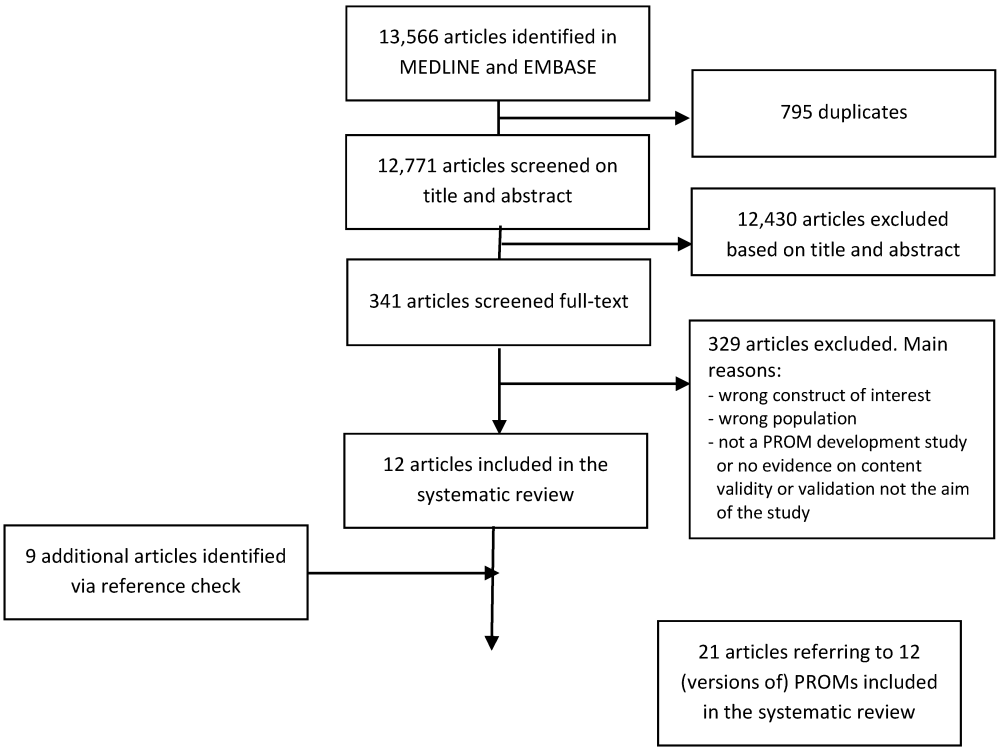
*Standard 1 in Box 3 in the COSMIN Risk of Bias checklist[32] was rated very good if CFA was performed, adequate if EFA was performed, doubtful if PCA was performed and inadequate if none of the previous was performed.

Appendix 4. Approach for grading the quality of the evidence

Grade factor	Downgrading	Definition
Risk of bias	0	Multiple studies of at least adequate quality *OR* one study of very good quality
	-1	Only one study of adequate quality *OR* multiple studies of doubtful quality
	-2	Only one study of doubtful quality *OR* multiple studies of inadequate quality
	-3	Only one study of inadequate quality
Imprecision (not for content validity, structural validity, and cross-cultural validity\ measurement invariance)	0	Total sample size of all studies >100
	-1	Total sample size of all studies 50-100
	-2	Total sample size of all studies <50
Inconsistency	0	Results are consistent *OR* results are summarized and rated per subset of studies, and subsequently graded
	-1	Overall rating based on the majority of consistent results
Indirectness	0	Does not occur; definitions for construct and/or target population have been stated in the inclusion criteria

0: high, -1: moderate, -2: low, -3: very low; Per protocol of the COSMIN guideline for systematic reviews: the quality of evidence for internal consistency cannot be higher than the quality of evidence for structural validity[23]

Appendix 5. Flowchart of the results from the search strategy



Appendix 6. Study populations involved in PROM development and content validity studies (marked with *)

PROM	Language	Country	Patients	Patient input	Mean age [range] yr	Gender % female	Disease characteristics	Professionals	Characteristics of professionals	Professional input
DFS[34]	English	UK	10 + 14 + 12	Concept elicitation and pilot testing relevance and comprehensibility	61 [46-74]	33-40%	Diabetic foot ulcers	0		
DFS-SF[35]	English	US	0					0		
DFS-SF[37]*	Chinese	China	6	Comprehensibility and relevance				0		
DFS-SF[39]*	Spanish	Spain	0					?		Evaluation content validity
PRO-DM-Thai[40]	Thai	Thailand	12 + 15	Concept elicitation/ comprehensibility, face validity	61 [49-70]	50%	T2DM, disease duration 11 [2-20] yr	9 + 17	3 physicians, 2 nurses, 2 pharmacists, and 2 nutritionists	Concept elicitation / relevance, comprehensiveness, comprehensibility
IWADL/ APPADL[41]	English	US	54 + 24	Concept elicitation, comprehensibility			T2DM and BMI of 25–40 kg/m2	0		
QOLID[43]	English/ Hindi[a]	India	20	Concept elicitation, comprehensibility			T2DM	8	4 clinicians and 4 diabetes educators	Relevance, comprehensibility
DQLCTQ[44]	English	US	30	Rating domains			T1DM (n=23), T2DM (n=7)	11	Clinicians / experts on HRQL research	Rating domains / evaluating face- and content validity of draft PROM
Diabetes-39[45]	English	US	?	Concept elicitation			Diabetes	?	Physicians, certified diabetes educators, pharmacists	Concept elicitation
Diabetes-39[46]*	Arabic	Jordan	30	Comprehensibility				0		
Diabetes-	Vietnamese	Vietnam	10	Comprehensibility				0		

39[47]*								
Diabetes-39[48]*	Portuguese	Brazil	4	Comprehensibility		0		
Diabetes-39 SF[51]	Chinese	Taiwan	0			0		
C-CWIS[54]	Chinese	China	20	Pilot testing	Diabetic foot ulcers	5	2 majored in medicine, 3 majored in nursing	Transcultural adjustment

a Language unsure

Appendix 7. Characteristics of included study populations for the assessment of measurement properties

	Population			**Disease characteristics**		**Instrument administration**			
PROM	**N**	**Mean age (SD) [range] yr**	**Gender % female**	**Disease**	**Disease duration mean (SD) yr**	**Setting**	**Country**	**Language**	**Response rate**
DFS[34]	48 + 54 + 71 = 173 Internal consistency: 48 Structural validity: 326 Known-groups: 102 (48+54) Responsiveness: 264	58 (11), 65 (12), 55 (15) N=326: ? N=264: ?	27%, 28%, 34% N=326: ? N=264: ?	T2DM and T1DM with neuropathic or mixed neuropathic/ ischemic foot ulcer	17 (11), 18 (12), 12 (9) N=326: ? N=264: ?	Diabetic foot centres	UK N=326: UK, US, Belgium, Denmark, France, Italy, Netherlands	English, and other	
DFS-SF study 1[35]	180	[27-87]	26%	Chronic, neuropathic, full-thickness, diabetic foot ulcers		Clinical trials	Belgium (16), Denmark (11), France (3), Germany (5), Italy (3), Netherlands (5), UK (243), US (40)	English, and other	
DFS-SF study 2[35]	252 + 95 = 347 Responsiveness: 252	[29-88], [33-83]	30%, 26%	Chronic, neuropathic, full-thickness, diabetic foot ulcers		Clinical trials	Austria (23), France (62), Germany (67), Greece (34), Italy (18), Netherlands (66), Switzerland (15) UK (16), US (95)	English, and other	
DFS-SF Polish[36]	212	63 (10)	28%	Diabetes (T2DM 86%)	18 (12)	Ambulatory setting of university	Poland	Polish	

						hospital			
DFS-SF Chinese[37]	60	71 (11)	40%	Current/healed diabetic foot ulcers	17 (10)	Outpatient/inpatient hospital	Hong Kong	Hong-Kong Chinese	
DFS-SF Greek[38]	110	60 (12)	39%	Diabetes (T2DM 86%) with diagnosed diabetic foot ulcer	17 (7)	General hospital	Greece	Greek	88%
DFS-SF Spanish[39]	141	68 (13)	33%	Diabetes (T2DM 95%) with new-onset diabetic foot ulcer		Diabetic foot unit	Spain	Spanish	
PRO-DM-Thai[40]	500 Construct validity: 200	66 (11) [32-90] N=200: 67 (11) [32-89]	67% N=200: 68%	T2DM (well-controlled and uncontrolled)	15 (8) N=200: 16 (8)	Outpatient clinic university hospital	Thailand	Thai	
IWADL/APPADL[41]	349	59 (9)	56%	T2DM with BMI 30-40 kg/m2		Internet panel	USA	English	
IWADL/APPADL[42]	106 Responsiveness: 40	52 (10) N=40: 52 (12)	69% N=40: 65%	T2DM with BMI at least 30 kg/m2		Weight loss centres, university-based weight loss program, or specialty clinic	USA	English	89%
QOLID[43]	150 Comparison instruments: 30 Known-groups: 210[a]	54 (7) [35-65] N=30 and N=210: similar to N=150	40% N=30 and N=210: similar to N=150	T2DM	11 (6) N=30 and N=210: similar to N=150	Diabetes centre	India	English/Hindi[b]	33% (150/460); 68% 30/44; 73% 30/41
DQLCTQ[44]	942 Internal	Type I diabetes: 34, type II	43% N=909: ?	T1DM (n=468) and T2DM (n=474)[c]	13 N=909: ?	Clinical trials	Canada (72), France (84), Germany (188),	English, German, French	

	consistency: 909 Reliability: 58 Responsiveness: 328	diabetes: 58 N=909: ? N=58: ? N=328: ?	N=58: ? N=328: ?		N=58: ? N=328: ?		US (598		
Diabetes-39 study 1[45]	516	52 (16)	55%	T1DM (33%) and T2DM (68%)	14 (11)	Diabetes care centre	US	English	52%
Diabetes-39 study 2[45]	165 + 262 = 427	62 (18), 55 (13)	55%, 65%	T1DM (20%, 10%) and T2DM (81%, 90%)	12 (11), 10 (8)	General practice, hospital diabetes clinic	US	English	70%, 41%
Diabetes-39 Arabic[46]	368	57 (10)	56%	T2DM	9 (7)	Outpatient clinic of university hospital	Jordan	Arabic	92%
Diabetes-39 Vietnamese[47]	286	20% <50, 45% 60-65, 36% >65	64%	T2DM	6 (5)	Tertiary hospital	Vietnam	Vietnamese	
Diabetes-39 Portuguese[48]	52	63 (9) [45-84]	65%	T2DM	9 (4)	Basic health service	Brazil	Portuguese	
Diabetes-39 Spanish[49]	249	5% <40, 61% 40-59, 35% ≥60, Males: 52.5 [24-75] years), Females: 55.7 [34-91]	63%	T2DM	9 (8)	Family medicine unit	Mexico	Spanish	96%
Diabetes-39 Taiwan[50]	280	63 (11)	47%	T1DM (1%) and T2DM (99%)	9 (6)	Outpatient clinic of teaching hospital	Taiwan	Chinese	
Diabetes-39 Taiwan[51]	265[d]			T1DM and T2DM		Diabetes clinics of a teaching	Taiwan	Chinese	

						hospital			
Diabetes-39 Thai[52]	397	58 (11)	74%	Diabetic people	6 (6)	Community hospital	Thailand	Thai	93%
Diabetes-39 German[53]	144 Reliability: 72 Responsiveness: 62-66	57 (8) N=72: ? N=62-66: ?	52% N=72: ? N=62-66: ?	T2DM	13 (10) N=72: ? N=62-66: ?	Special hospital/outpatient clinic for people with diabetes	Germany	German	85%
Diabetes-39 SF[51]	265[d]			T1DM and T2DM		Diabetes clinics of a teaching hospital	Taiwan	Chinese	
C-CWIS[54]	131 Internal consistency: 20 & 131	68 (11) N=20: ?	34% N=20: ?	T2DM with diabetic foot ulcers	14 (8) N=20: ?	Outpatient/inpatient diabetic foot of an integrated hospital	China	Chinese	

a N=60 for HbA1C values; b Language unsure; c Analyses conducted separately for each group; pooled analyses performed because results were similar; d Sample from study of Huang[50]

Appendix 8. Information on feasibility of PROMs

PROM	Type and ease of administration	Length of instrument[a]	Response options included subscale	Completion time	Patient's required mental and physical ability level	Ease of score calculation	Copyright
DFS[34]	Self-report	11 subscales, 58 items: Leisure (5); Physical health (6); **Daily activities (6)**; Emotions (17); Noncompliance (2); Family (5); Friends (5); Treatment (4); Satisfaction (1); Positive attitude (5); Financial (2)	Daily activities: 1 = none of the time, 2 = a little bit of the time, 3 = some of the time, 4 = most of the time, and 5 = all of the time			Scores are based on the sum of items associated with a subscale if at least 50% of the items in a scale are completed. When necessary, raw item scores are reverse coded so that the minimum possible score (1) represents the worst quality of life, and the maximum possible score (5) represents the best quality of life (all items except in the positive attitude subscale). Each subscale is scored from 0 to 100, higher scores indicate better quality of life.	Johnson & Johnson Research & Development, LCC
DFS-SF[35-39]	Self-report/interview-based	6 subscales, 29 items: Leisure (5); **dependence/ daily life (5)**; negative emotions (6); physical health (5); worried about ulcers/feet (4); bothered by ulcer care (4)	Dependence/ daily life: 1 = none of the time, 2 = a little of the time, 3 = some of the time, 4 = most of the time, and 5 = all of the time	12.5 minutes for interview-based administration		Scores are based on the sum of items associated with a subscale if at least 50% of the items in a scale are completed. In case of item-level missing data <50%, the subscale score is calculated by substituting the mean item score for the missing item values. Raw item scores are reverse coded so that	Johnson & Johnson Research & Development, LCC

						the minimum possible score (1) represents the worst quality of life, and the maximum possible score (5) represents the best quality of life. Each subscale is scored from 0 to 100, higher scores indicate better quality of life.	
PRO-DM-Thai[40]	Self-report/interview-based	7 subscales, 44 items: **Physical function (5);** symptoms (7); Psychological wellbeing (5); Self-care management (12); Social wellbeing (5); Global judgements of health (5); Satisfaction with care and flexibility of treatment (5)	Unknown, PROM could not be retrieved	30 minutes		Not reported.	
IWADL/APPADL[41, 42]	Self-report	1 subscale, 7 items: **Physical activities of daily living (7)**	1-5: 1 = unable to do, 5 = not at all difficult	<5 minutes	Flesch Kincaid reading level: 9th grade	Total scores are derived by adding item scores (minimum = 1, maximum = 5) and then dividing by the number of items, so that the minimum and maximum total scores are 1 and 5, respectively. Total scores can be transformed to 0-100. Higher scores correspond to greater ability to do physical daily activities.	Publicly available
QOLID[43]	Interview-based	8 subscales, 34 items: Role limitations due to physical health (social life, work, traveling) (6);	1-5: 1 referring to poorest outcome, 5 to best outcome	Mean: 7.8 minutes, SD: 2.8 minutes		A score for each domain is calculated by adding items' scores after mean imputation for 'not	

		Physical endurance (6); General health (3); Treatment satisfaction (4); Symptom botherness (3); Financial worries (4); Emotional/ mental health (5); Diet advise tolerance (3)			applicable' values. Each domain score is standardized by dividing by the maximum possible domain score and multiplying by 100. All domain scores are added and divided by 8 (the number of domains) to obtain an overall score. Standardized scores range 0-100.	
DQLCTQ[44]	Self-report	8 subscales, 57 items: **Physical function (6)**; Energy/fatigue (5); Health distress (6); Mental health (5); Satisfaction (DQOL – 18, excl. 3 skip pattern questions); Treatment satisfaction (3); Treatment flexibility (10); Frequency of symptoms (7)	Physical function: 1 = limited for more than four weeks, 2 = limited for four weeks or less, 3 = not limited at all	10 minutes	The average of a domain is computed by summing up scores within the domain, and dividing the sum by the number of items in the domain. If 50% or more of the items are missing, the average score should not be calculated and the domain score is treated as missing. Domain scores are converted to a 100-point scale. Higher scores indicate better quality of life.	Yes
Diabetes-39[44-52]	Self-report/interview-based	5 subscales, 39 items: **Energy and mobility (15)**; Diabetes control (12); Anxiety and worry (4); Social burden (5); Sexual functioning (3)[b]	English/Arabic: VAS marked 1-7: 1 = not affected at all, 7 = extremely affected Other: 1-7: 1 = not affected at all, 7 = extremely affected	10-15 minutes	English/Arabic: Respondents place an 'X' on a modified visual analogue scale ranging from 1 (= not affected at all) to 7 (= extremely affected). The response is measured to the nearest quarter of a centimeter. Any response falling between two of the quarter graduations is rounded to the higher	

				quarter of a centimeter. If more than 7 items are missing, the questionnaire is not analyzed. In case of 7 or less missing items, the modal response within each subscale served as a proxy for missing data. Each scale score is transformed to 0-100, with 0 indicating the least impact on quality of life and 100 indicating the most impact. Other: Respondents place an 'X' in one of the boxes numbered 1 to 7, which are on a horizontal bar. The number marked, without any 0.5 point approximation, for each subscale is summed, and then transported to a scale from 0 to 100, with a higher score indicating greater impact on quality of life.[c]
Diabetes-39 SF[51]	Interview-based	5 subscales, 22 items: **Energy and mobility (5)**; Diabetes control (5); Anxiety and worry (4); Social burden (5); Sexual functioning (3)	1-7: 1 = not affected at all, 7 = extremely affected	Scores are given on a 7-point Likert scale. Subscale scores are calculated by summing all responses, high scores represent poor quality of life.
C-CWIS[54]	Self-report/interview-based	3 subscales, 25 items: **Physical symptoms and everyday living (5)**; Social life (7); Well-being (6)	1-5: 1 = not at all, 5= always	Total item scores includes patient's perception of the experience and the associated stress. To calculate scale scores, the

		item scores and number of sub-evaluations are summated for each scale (unclear how this is exactly done, validated formula is used), subscale scores range from 0-100, higher scores indicate better health-related quality of life.

a Bold subscales measure physical functioning; b Thai version: 6 subscales and 39 items: Energy and mobility (10), Diabetes control (13), Anxiety and worry (4), Social burden (6), Sexual functioning (3), Other health problems and diabetic complications (3); German version: 5 subscales, 39 items: Diabetes and treatment (7), Physical impairment (7), Social stress (5), Physical illness (5), Sexual problems (3), subscale unknown for remaining 12 items; c Vietnamese version: If more than 4 items are missing (except from the sexual functioning domain), the questionnaire is not analyzed. For the energy and mobility scale, of more than 3 items are missing, a scale score is not calculated. If 3 of less items are missing, the missing value is replaced by the mean scale score

Appendix 9. Information on interpretability of PROMs

PROM – *subscale*	Distribution of scores in the study population	Percentage of missing items or percentage of missing scores	Floor and ceiling effects	Scores and change scores available for relevant (sub)groups	Minimal important change (MIC) or minimal important difference (MID)
DFS – *Daily activities*[34]				Healed ulcer: ~69, Current ulcer: ~63[a]	
DFS-SF – *Dependence/ daily life*[35-39]	[36]: mean=47.7, median =50.0, SD=29.3 [37]: mean=71.4, median =85.0, SD=32.9 [38]: mean=56.3, median=55.0, SD=25.7	[36]: 0.0-1.5% [37]: 0%	[36]: 7.8% floor, 2.9% ceiling [37]: 5% floor, 30% ceiling [38]: 0.9% floor, 5.5% ceiling	[35]: Pre vs. post closure of target ulcer change score – study 1: +3.7; study 2 +10.0 [37]: Healed vs. unhealed ulcer Change score: +13.9 [38]: >1 complication: 44.7, 1 complication: 55.8, No complication: 69.5	
PRO-DM-Thai – *Physical function*[40]					
IWADL/APPADL – *(Physical) activities of daily living*[41, 42]	[41]: Mean=3.3, SD=1.1 [42]: Mean=3.3, SD=1.0		[41]: 4-31% floor per item, 8-32% ceiling per item [42]: 6% floor effect, 11% ceiling effect		[42]: If transformed to 0-100: SEM = 6.3; MIC based on weight loss: 13.6; MIC based on ability to perform daily physical activities: 9.8
QOLID – *Physical endurance*[43]	Mean=25.3, SD=5.5 Standardized mean =84.3, SD=18.4	20.7%		HbA1c ≤8: 87.1, HbA1c >8: 81.8; Insulin: 81.6, Non-insulin: 86.1; Comorbidity ≤1: 77.9, Comorbidity >1: 88.6; Male: 89.4, Female: 76.4	
DQLCTQ – *Physical function*[44]				HbA1c tight: 89.3, HbA1c poor: 85.1; Type 1: 94.7, Type 2: 77.9; Male: 87.9, Female: 84.1; Good control: 88.2, Poor control: 81.3	
Diabetes-39 – *Energy and mobility (pilot version - 14*		Study 1 - Total questionnaire: 0.3-0.45%			

items)[45]				
Diabetes-39 – *Energy and mobility (15 items)*[45-51]	[46]: mean=50.5, SD=21.1 [47]: median=41.1, 25th percentile =22.2, 75th percentile =60.0, range=0.0-91.1 [48]: median=51.5, mean=48.8, SD=14.5, range (possible: 15-105) =20-83 [49]: median=30, 25th percentile =16, 75th percentile =50	[45]: Study 2 - Total questionnaire: 0.3-0.45%	[46]: 0.3% ceiling, 0.8% floor [48]: 1.9 % floor, 1.9% ceiling	[47]: Male: 40.0, Female: 43.3; Insulin: 46.1, Non-insulin: 40.0; Comorbidity: 43.3, No comorbidity: 26.1 Complication: 49.4, No complication: 37.8 [49]: Male: 27, Female: 35
Diabetes-39 Thai – *Energy and mobility (10 items)*[52]		Total questionnaire: <2%		Comorbidity: 28.8, No comorbidity: 26.0 Insulin: 35.8, Non-insulin: 26.6; Complications: 37.0, No complications: 27.1
Diabetes-39 German – *Physical impairment (7 items)*[53]		N=1-4 for all subscales		Insulin: 18.4, No insulin: 15.8; ≤1 Complication: 11.5, >1 Complication: 20.4
Diabetes-39 SF – *Energy and mobility (5 items)*[51]				
C-CWIS – *Physical symptoms and everyday living*[54]	Item score ranges: mean=4.98-8.78, SD=1.79-2.58			

a Read from histogram

Appendix 10. Extensive results of studies on measurement properties

PROM – *subscale*	**Country (language) in which the PROM was evaluated**	**Structural validity**			**Internal consistency**			**Cross-cultural validity\ measurement invariance**			**Reliability**		
		n	**Meth qual**	**Result (rating)**	**n**	**Meth qual**	**Result (rating)**	**n**	**Meth qual**	**Result (rating)**	**n**	**Meth qual**	**Result (rating)**
DFS – *Daily activities*[34]	English, and other	326	Inade quate	Daily life (3 items) (?) and dependence (4 items) (?)	48	Inade quate	Daily activities: α^{a}=0.85 (?)				?	Inadequa te	Daily life: r=0.68 (-) Dependence: r=0.84 (+)
Pooled or summary result (overall rating)		**326**		**Daily life (3 items) and dependence (4 items) (?)**	**48**		**Daily activities: α=0.85 (?)**				**?**	**Very low**	**Daily life: r=0.68 (-) Dependence: r=0.84 (+)**
DFS-SF study 1 – *Dependence /daily life*[35]	English, and other	180	Inade quate	Dependence/ daily life (5 items): factor loadings 0.53-0.85; one cross- loader in whole instrument (?)	180	Very good	Dependence /daily life: α=0.88 – 0.91[b] (+)				?	Inadequa te	Dependence /daily life: ICC=0.77 (+)
DFS-SF study 2 – *Dependence /daily life*[35]	English, and other	347	Very good	Dependence/ daily life (5 items): non-normed fit index = 0.93; CFI = 0.94; incremental fit index = 0.94; goodness of fit index = 0.87; RMSEA = 0.058 (+)	347	Very good	Dependence /daily life: α=0.85 - 0.88[c] (+)				?	Inadequa te	Dependence /daily life: ICC=0.74 (+)
DFS-SF Polish – *Dependence*	Polish				212	Very good	Dependence /daily life:	212	Inadeq uate	DIF in item 3c			

/daily life[36]							α=0.90 (+)			for sex, (p-value Chi2 = 0.02) 6 variables were tested in 5 items (+)			
DFS-SF Chinese – *Dependence /daily life*[37]	Hong Kong Chinese				60	Very good	Dependence /daily life: α=0.89 (+)						
DFS-SF Greek – *Dependence /daily life*[38]	Greek				110	Very good	Dependence /daily life: α=0.87 (+)						
DFS-SF Spanish – *Dependence /daily life*[39]	Spanish	141	Inade quate	Dependence/ daily life (5 items): CFA: CFI = 0.844; RMSEA = 0.095; SRMR = 0.093 (-) EFA: 65.5% total variance explained; dependence/ daily life scale: factor loadings 0.58-1.00 (+)	141	Very good	Dependence /daily life: α=0.87 (+)				141	Doubtful	Dependence /daily life: ICC cons=0.77 (+)
Pooled or summary result (overall rating)		**347**	**High**	**Dependence/ daily life (5 items) (+)**	**105 0**	**High**	**Dependence /daily life: α=0.85-0.91**	**212**	**Very low**	**DIF for sex in one item**	**>141**	**Low**	**Dependence /daily life: ICC=0.74-0.77 (+)**

							(+)			**(+)**			
PRO-DM-Thai – *Physical function*[40]	Thai	500	Very good	Physical function (5 items): Goodness-of-Fit index 0.998; Adjusted Goodness-of-Fit index 0.991; RMSEA: 0.000 (+)	500	Very good	Physical function: α=0.82 (+)						
Pooled or summary result (overall rating)		**500**	**High**	**Physical function (5 items) (+)**	**500**	**High**	**Physical function: α=0.82 (+)**						
IWADL/APPADL – *(Physical) activities of daily living*[41]	English	349	Doubtful	(physical) activities of daily living (7 items): 73% variance explained; factor loadings 0.82-0.90; eigenvalue 5.1 (+)	349	Very good	(physical) activities of daily living: α=0.94 (+)						
IWADL/APPADL – *(Physical) activities of daily living*[42]	English				106	Very good	(physical) activities of daily living: α≥0.89[d] (+)				106	Adequate	(physical) activities of daily living: ICC agr=0.91 (+)
Pooled or summary result (overall rating)		**349**	**Low**	**(physical) activities of daily living (7 items) (+)**	**455**	**Low[e]**	**(physical) activities of daily living: α≥0.89 (+)**				**106**	**Moderate**	**(physical) activities of daily living: ICC agr=0.91 (+)**
QOLID – *Physical*	English/Hindi[f]	150	Inadequate	Physical endurance (6	150	Very good	Physical endurance:						

endurance[43]				items): 5.9%variance explained (total 49.9%); factor loadings 0.52 – 0.72 (+)			α=0.85 (?)						
Pooled or summary result (overall rating)		**150**	**Very low**	**Physical endurance (6 items): (+)**	**150**		**Physical endurance: α=0.85 (?)**						
DQLCTQ – *Physical function*[44]	English, German, French				909	Doubtful	Physical function: α=0.85 (?)				58	Adequate	Physical function: ICC=0.83 (+)
Pooled or summary result (overall rating)					**909**		**Physical function: α=0.85 (?)**				**58**	**Low**	**Physical function: ICC=0.83 (+)**
Diabetes-39 – *Energy and mobility (pilot version - 14 items)*[45]	English	516	Adequate	Energy and mobility (14 items): Seven factors with eigenvalue >1; 77% total variance explained; item loadings >0.50 (+)	516	Very good	All subscales: α=0.81-0.92 (+)						
Pooled or summary result (overall rating)		**516**	**Moderate**	**Energy and mobility (14 items) (+)**	**516**	**Moderate[e]**	**All subscales: α=0.81-0.92 (+)**						
Diabetes-39 – *Energy and mobility (15 items)*[45]	English	427	Adequate	Energy & mobility (15 items): Five factors with eigenvalue >1;	427	Very good	Energy and mobility: α=0.93 (+)						

				90% total variance explained; items loadings >0.50 (+)									
Diabetes-39 Arabic – *Energy and mobility (15 items)*[46]	Arabic	368	Doubtful	Energy & mobility (15 items): 83% total variance explained; item-scale correlations >0.40 (+)	368	Very good	Energy and mobility: α=0.89 (+)						
Diabetes-39 Vietnamese – *Energy and mobility (15 items)*[47]	Vietnamese				286	Very good	Energy and mobility: α=0.92 (+)				286[g]	Doubtful	Energy and mobility: ICC=0.91 (+)
Diabetes-39 Portuguese – *Energy and mobility (15 items)*[48]	Portuguese				52	Very good	Energy and mobility: α=0.79 (+)						
Diabetes-39 Spanish – *Energy and mobility (15 items)*[49]	Spanish				249	Very good	Energy and mobility: α=0.92 (+)				249	Doubtful	Energy and mobility: spearman r=0.84 (+)
Diabetes-39 Taiwan – *Energy and mobility (15 items)*[50]	Chinese				280	Very good	All subscales: α=0.82-0.93 (+)						
Diabetes-39 Taiwan –	Chinese	265	Adequate	Energy and mobility (15									

Energy and mobility (15 items)[51]				items): CFI: 0.92; TLI: 0.98; RMSEA: 0.085 (+)									
Pooled or summary result (overall rating)		**1060**	**High**	**Energy and mobility (15 items) (+)**	**1662**	**High**	**Energy and mobility: α=0.79-0.93 (+)**				**535**	**Moderate**	**Energy and mobility: spearman r=0.84, ICC=0.91 (+)**
Diabetes-39 Thai – *Energy and mobility (10 items)*[52]	Thai	397	Adequate	Energy and mobility (10 items): Factor loadings 0.28-0.80; 1 cross-loader in scale; total variance explained 62.1% (-)	397	Very good	Energy and mobility: α=0.94 (?)						
Pooled or summary result (overall rating)		**397**	**Moderate**	**Energy and mobility (10 items) (-)**	**397**		**Energy and mobility: α=0.94 (?)**						
Diabetes-39 German – *Physical impairment (7 items)*[53]	German	144	Inadequate	Physical impairment (7 items): Factor loadings >0.50 (?)	144	Doubtful	Physical impairment: α=0.84 (?)				72[6]	Doubtful	Physical impairment: 0.88[h] (?)
Pooled or summary result (overall rating)		**144**		**Physical impairment (7 items) (?)**	**144**		**Physical impairment: α=0.84 (?)**				**72**		**Physical impairment: 0.88 (?)**
Diabetes-39 SF – *Energy and mobility (5 items)*[51]	Chinese	265	Very good	Energy and mobility (5 items): CFI: 0.966; TLI: 0.993; RMSEA: 0.058 (+)									
Pooled or summary result		**265**	**High**	**Energy and**									

(overall rating)				**mobility (5 items) (+)**									
C-CWIS – *Physical symptoms and everyday living*[54]	Chinese	131	Adequate	Physical symptoms and everyday living (12 items): Factor loadings 0.53-0.78 (one item with loading 0.16 loaded on a different factor but was retained in its original factor); 39.4% variance explained (total 57.2%); eigen value 9.9 (+)	Pilot: 20 Formal: 131	Very good	Physical symptoms and everyday living: Pilot: α=0.73 (+) Formal: α=0.92 (+)						
Pooled or summary result (overall rating)		**131**	**Moderate**	**Physical symptoms and everyday living (12 items): (+)**	**151**	**Moderate[e]**	**Physical symptoms and everyday living: α=0.73-0.92 (+)**						

Appendix 10. Continued

PROM – *subscale*	**Country (language) in which the PROM was evaluated**	**Measurement error**			**Criterion validity**			**Hypotheses testing for construct validity** **a=comparison with other instruments** **b=comparison between subgroups**			**Responsiveness** **a=comparison to gold standard** **b=comparison with other instruments** **c=comparison between subgroups** **d=before and after intervention**		
		n	**Meth**	**Result (rating)**	**n**	**Meth**	**Result (rating)**	**n**	**Meth**	**Result**	**n**	**Meth**	**Result (rating)**

			qual			**qual**			**qual**	**(rating)**		**qual**	
DFS – *Daily activities*[34]	English, and other							a. 173 b. 102[i]	a. Adequate b. Very good	a. Results in line with 4 hypos (4+); results not in line with 1 hypo (1-) b. No data provided (?)	c. 264	c. Doubtful	c. Results not in line with 2 hypos (2-)
Pooled or summary result (overall rating)								**a. 173 b. 102**	**a. Moderate**	**a. 4+ and 1- (+) b. (?)**	**c. 264**	**c. Low**	**c. 2- (-)**
DFS-SF study 1 – *Dependence /daily life*[35]	English, and other							a. 180	a. Adequate	a. Results in line with 2 hypos (2+); results not in line with 3 hypos (3-)	c. 180	c. Very good	c. Results not in line with 1 hypo (1-)
DFS-SF study 2 – *Dependence /daily life*[35]	English, and other							a. 347	a. Adequate	a. Results in line with 2 hypos (2+); results not in	c. 252	c. Very good	c. Results in line with 1 hypo (1+)

										line with 3 hypos (3-)			
DFS-SF Polish – *Dependence /daily life*[36]	Polish							a. 212 b1. 212 b2. 212	a. Very good b1. Inadequate b2. Doubtful	a. Results in line with 4 hypos (4+); results not in line with 3 hypos (3-) b1. Results in line with 1 hypo (1+); results not in line with 3 hypos (3-) b2. No data provided (?)			
DFS-SF Chinese – *Dependence /daily life*[37]	Hong Kong Chinese							a. 60 b. 60	a. Very good b. Very good	a. Results in line with 5 hypos (5+); results not in line with			

										2 hypos (2-) b. Results in line with 1 hypo (1+)[8]; results not in line with 1 hypo (1-)			
DFS-SF Greek – *Dependence /daily life*[38]	Greek							a. 110 b. 110[j]	a. Very good b. Very good	a. Results in line with 4 hypos (4+); results not in line with 3 hypos (3-) b. Results in line with 14 hypos (14+); results not in line with 1 hypo (1-)			
DFS-SF Spanish –	Spanish							a. 141	a. Adequa	a. Results in line	c. 141	c. Very good	c. Results in line with 1 hypo (1+)

Dependence /daily life[39]									te	with 4 hypos (4+); results not in line with 1 hypo (1-)			
Pooled or summary result (overall rating)								**a. 1050 b. 170**	**a. High b. High**	**a. 21+ and 15- (±) b. 15+ and 2- (+)**	**c. 573**	**c. High**	**c. 2+ and 1- (±)**
PRO-DM-Thai – *Physical function*[40]	Thai							b1. 200 b2. 200	b1. Inadequate b2. Very good	b1. Results not in line with 1 hypo (1-) b2. No data provided (?)			
Pooled or summary result (overall rating)								**b. 200**	**b. Very low**	**b. 1- (-)**			
IWADL/APPADL – *(Physical) activities of daily living*[41]	English							b. 349	b. Very good	b. Results in line with 12 hypos (12+); results not in line with 22 hypos			

										(22-)			
IWADL/APPADL – *(Physical) activities of daily living*[42]	English	106	Doubtful	(physical) activities of daily living: SEM=6.3; SDC=17.5; MIC=9.8-13.6[k] (-)							d. 40	d. Very good	d. Results in line with 2 hypos (2+); results not in line with 1 hypo (1-)
Pooled or summary result (overall rating)		**106**	**Low**	**(physical) activities of daily living: SEM=6.3; SDC=17.5; MIC=9.8-13.6[k] (-)**				**b. 349**	**b. High**	**b. 12+ and 22- (±)**	**d. 40**	**d. Low**	**d. 2+ and 1- (±)**
QOLID – *Physical endurance*[43]	English/Hindi[f]							a. 30 b. 210[l]	a. Adequate b. Very good	a. Results in line with 3 hypos (3+); results not in line with 3 hypos (3-) b. Results not in line with 3 hypos (3-)			
Pooled or summary result (overall rating)								**a. 30 b. 210**	**a. Very low b. High**	**a. 3+ and 3 - (±) b. 3- (-)**			

DQLCTQ – *Physical function*[44]	English, German, French							b. 942 [m]	b. Very good	b. Results in line with 1 hypo (1+); results not in line with 1 hypo (1-)	c. 328	c. Very good	c. Results not in line with 2 hypos (2-)
Pooled or summary result (overall rating)								**b. 942**	**b. High**	**b. 1+ and 1- (±)**	**c. 328**	**c. High**	**c. 2- (-)**
Diabetes-39 – *Energy and mobility (pilot version - 14 items)*[45]	English							b. 516	b. Adequate	b. No data provided (?)			
Pooled or summary result (overall rating)								**b. 516**		**b. (?)**			
Diabetes-39 – *Energy and mobility (15 items)*[45]	English							a. 427 b. 427	a. Adequate b. Doubtful	a. Results in line with 6 hypos (6+); results not in line with			

										4 hypos (4-) b. No data provided (?)			
Diabetes-39 Arabic – *Energy and mobility (15 items)*[46]	Arabic							a. 368	a. Adequate	a. Results in line with 2 hypos (2+); results not in line with 3 hypos (3-)			
Diabetes-39 Vietnamese – *Energy and mobility (15 items)*[47]	Vietnamese							b. 286	b. Very good	b. Results in line with 4 hypos (4+); results not in line with 6 hypos (6-)[12]			
Diabetes-39 Portuguese – *Energy and mobility (15 items)*[48]	Portuguese							b. 52	b. Very good	b. No data provided (?)			
Diabetes-39 Spanish – *Energy and mobility (15*	Spanish							b. 249	b. Very good	b. No data provided (?)			

items)[49]													
Diabetes-39 Taiwan – *Energy and mobility (15 items)*[50]	Chinese							a. 280 b. 280	a. Very good b. Very good	a. No hypos defined (?) b. Results in line with 7 hypos (7+); results not in line with 5 hypos (5-)			
Diabetes-39 Taiwan – *Energy and mobility (15 items)*[51]	Chinese												
Pooled or summary result (overall rating)								**a. 795** **b. 566**	**a. High** **b. High**	**a. 8+ and 7- (±)** **b. 11+ and 11- (±)**			
Diabetes-39 Thai – *Energy and mobility (10 items)*[52]	Thai							a. 397 b. 397	a. Adequate b. Very good	a. Results in line with 2 hypos (2+); results not in line with 3 hypos			

										(3-) b. Results in line with 4 hypos (4+); results not in line with 2 hypo (2-)			
Pooled or summary result (overall rating)								**a. 397 b. 397**	**a. Moderate b. High**	**a. 2+ and 3- (±) b. 4+ and 2- (±)**			
Diabetes-39 German – *Physical impairment (7 items)*[53]	German							a. 144 b. 144	a. Very good b. Very good	a. No hypos defined (?) b. Results in line with 4 hypos (4+)	d. 62-66	d. Doubtful	d. No data provided (?)
Pooled or summary result (overall rating)								**a. 144 b. 144**	**b. High**	**a. (?) b. 4+ (+)**	**d. 62-66**		**d. (?)**
Diabetes-39 SF – *Energy and mobility (5 items)*[51]	Chinese												

Pooled or summary result (overall rating)													
C-CWIS – *Physical symptoms and everyday living*[54]	Chinese							a. 131 b. 131	a. Inadequate b. Inadequate	a. No data provided (?) b. Result in line with 1 hypo (1+)			
Pooled or summary result (overall rating)									**b. Very low**	**a. (?) b. 1+ (+)**			

a α refers to Cronbach's alpha; b Four moments of measurement; c Three moments of measurement; d Three moments of measurements; T1: n=119, T2: n=106, T3: n=40; e Per protocol of the COSMIN guideline for systematic reviews: the quality of evidence for internal consistency cannot be higher than the quality of evidence for structural validity[23]

f Language unsure; g Sample size unsure; h Reliability parameter unknown; i One of the known-groups tested in the hypotheses was small (n=10); j Some of the known-groups tested in the hypotheses were small (n<10); k MIC based on ability to perform daily physical activities and MIC based on weight loss, respectively; 10 n=60 for HbA1C levels; l n=274 for HbA1C levels; m One of the known-groups tested in the hypotheses was small (n=14)

Appendix 11. Extensive summary of findings

STRUCTURAL VALIDITY			
PROM – *subscale*	**Summary or pooled result**	**Overall rating**	**Quality of evidence**
DFS – *Daily activities*[34]	Two subscales: Daily life (3 items) and Dependence (4 items)	Indeterminate	No level of evidence: rating is indeterminate
DFS-SF – *Dependence/daily life*[35-39]	Unidimensional scale	Sufficient	High: one very good study (results of inadequate study are ignored)
PRO-DM-Thai – *Physical function*[40]	Unidimensional scale	Sufficient	High: one very good study
IWADL/ APPADL – *(Physical) activities of daily living*[41, 42]	Unidimensional scale	Sufficient	Low: one doubtful study
QOLID – *Physical endurance*[43]	Unidimensional scale	Sufficient	Very low: one inadequate study
DQLCTQ – *Physical function*[44]			
Diabetes-39 – *Energy and mobility (pilot version - 14 items)*[45]	Unidimensional scale	Sufficient	Moderate: one adequate study
Diabetes-39 – *Energy and mobility (15 items)*[45-51]	Unidimensional scale	Sufficient	High: two adequate studies
Diabetes-39 Thai – *Energy and mobility (10 items)*[52]	Structural validity unconfirmed	Insufficient	Moderate: one adequate study
Diabetes-39 German – *Physical impairment (7 items)*[53]	Structural validity unconfirmed	Indeterminate	No level of evidence: rating is indeterminate
Diabetes-39 SF – *Energy and mobility (5 items)*[51]	Unidimensional scale	Sufficient	High: one very good study
C-CWIS – *Physical symptoms and everyday living*[54]	Unidimensional scale	Sufficient	Moderate: one adequate study

INTERNAL CONSISTENCY			
PROM – *subscale*	**Summary or pooled result**	**Overall rating**	**Quality of evidence**
DFS – *Daily activities*[34]	Cronbach's alpha: 0.85; n=48	Indeterminate	No level of evidence: rating is indeterminate because structural validity is

			indeterminate
DFS-SF – *Dependence/daily life*[35-39]	Cronbach's alpha: 0.85-0.91; consistent results; n=1050	Sufficient	High: six very good studies
PRO-DM-Thai – *Physical function*[40]	Cronbach's alpha: 0.82; n=500	Sufficient	High: one very good study
IWADL/ APPADL – *(Physical) activities of daily living*[41, 42]	Cronbach's alpha: ≥0.89; consistent results; n=455;	Sufficient	Low: two very good studies, but structural validity is low
QOLID – *Physical endurance*[43]	Cronbach's alpha: 0.85; n=150	Indeterminate	No level of evidence: rating is indeterminate because structural validity is very low
DQLCTQ – *Physical function*[44]	Cronbach's alpha: 0.85; n=909	Indeterminate	No level of evidence: rating is indeterminate because structural validity is not assessed
Diabetes-39 – *Energy and mobility (pilot version - 14 items)*[45]	Cronbach's alpha: 0.81-0.92; n=516	Sufficient	Moderate: one very good study, but structural validity is moderate
Diabetes-39 – *Energy and mobility (15 items)*[45-51]	Cronbach's alpha: 0.79-0.93; consistent results; n=1622	Sufficient	High: six very good studies
Diabetes-39 Thai – *Energy and mobility (10 items)*[52]	Cronbach's alpha: 0.94; n=397	Indeterminate	No level of evidence: rating is indeterminate because structural validity is insufficient
Diabetes-39 German – *Physical impairment (7 items)*[53]	Cronbach's alpha: 0.84; n=144	Indeterminate	No level of evidence: rating is indeterminate because structural validity is indeterminate
Diabetes-39 SF – *Energy and mobility (5 items)*[51]			
C-CWIS – *Physical symptoms and everyday living*[54]	Cronbach's alpha: 0.73-0.92; n=151	Sufficient	Moderate: one very good study, but structural validity is moderate

CROSS-CULTURAL VALIDITY\MEASUREMENT INVARIANCE			
PROM – *subscale*	**Summary or pooled result**	**Overall rating**	**Quality of evidence**

DFS – *Daily activities*[34]			
DFS-SF – *Dependence/daily life*[35-39]	DIF for gender in one item, while five items were tested for six variables; n=212	Sufficient	Very low: one inadequate study
PRO-DM-Thai – *Physical function*[40]			
IWADL/ APPADL – *(Physical) activities of daily living*[41, 42]			
QOLID – *Physical endurance*[43]			
DQLCTQ – *Physical function*[44]			
Diabetes-39 – *Energy and mobility (pilot version - 14 items)*[45]			
Diabetes-39 – *Energy and mobility (15 items)*[45-51]			
Diabetes-39 Thai – *Energy and mobility (10 items)*[52]			
Diabetes-39 German – *Physical impairment (7 items)*[53]			
Diabetes-39 SF – *Energy and mobility (5 items)*[51]			
C-CWIS – *Physical symptoms and everyday living*[54]			

RELIABILITY			
PROM – *subscale*	**Summary or pooled result**	**Overall rating**	**Quality of evidence**
DFS – *Daily activities*[34]	Daily life: r=0.68; n=? Dependence: r=0.84; n=?	Daily life: insufficient Dependence: sufficient	Very low: one inadequate study
DFS-SF – *Dependence/daily life*[35-39]	ICC: 0.74-0.77; consistent results; n>141	Sufficient	Low: one doubtful study
PRO-DM-Thai – *Physical function*[40]			
IWADL/ APPADL – *(Physical) activities of daily living*[41, 42]	ICC: 0.91; n=106	Sufficient	Moderate: one adequate study

QOLID – *Physical endurance*[43]	ICC: 0.83; n=58	Sufficient	Low: one adequate study, sample size 58
DQLCTQ – *Physical function*[44]			
Diabetes-39 – *Energy and mobility (pilot version - 14 items)*[45]			
Diabetes-39 – *Energy and mobility (15 items)*[45-51]	ICC: 0.91, spearman r: 0.84; consistent results; n=535	Sufficient	Moderate: two doubtful studies
Diabetes-39 Thai – *Energy and mobility (10 items)*[52]			
Diabetes-39 German – *Physical impairment (7 items)*[53]	0.88; n=72	Indeterminate	No level of evidence: rating is indeterminate
Diabetes-39 SF – *Energy and mobility (5 items)*[51]			
C-CWIS – *Physical symptoms and everyday living*[54]			

MEASUREMENT ERROR			
PROM – *subscale*	**Summary or pooled result**	**Overall rating**	**Quality of evidence**
DFS – *Daily activities*[34]			
DFS-SF – *Dependence/daily life*[35-39]			
PRO-DM-Thai – *Physical function*[40]			
IWADL/ APPADL – *(Physical) activities of daily living*[41, 42]	SEM: 6.3, SDC: 17.5, MIC: 9.8-13.6; n=106	Insufficient	Low: one doubtful study
QOLID – *Physical endurance*[43]			
DQLCTQ – *Physical function*[44]			
Diabetes-39 – *Energy and mobility (pilot version - 14 items)*[45]			
Diabetes-39 – *Energy and mobility (15 items)*[45-51]			
Diabetes-39 Thai – *Energy and mobility (10 items)*[52]			

Diabetes-39 German – *Physical impairment (7 items)*[53]			
Diabetes-39 SF – *Energy and mobility (5 items)*[51]			
C-CWIS – *Physical symptoms and everyday living*[54]			

HYPOTHESES TESTING FOR CONSTRUCT VALIDITY **a=comparison with other instruments** **b=comparison between subgroups**			
PROM – *subscale*	**Summary or pooled result**	**Overall rating**	**Quality of evidence**
DFS – *Daily activities*[34]	a. 4 out of 5 hypotheses confirmed; n=173 b. No data provided	a. Sufficient b. Indeterminate	a. Moderate: one adequate study b. No level of evidence: rating is indeterminate
DFS-SF – *Dependence/daily life*[35-39]	a. 21 out of 36 hypotheses confirmed; n=1050 b. 15 out of 17 hypotheses confirmed; n=170	a. Inconsistent b. Sufficient	a. High: three very good studies b. High: two very good studies (results of inadequate study and study with no data provided are ignored)
PRO-DM-Thai – *Physical function*[40]	b. 0 out of 1 hypotheses confirmed; n=200	b. Insufficient	b. Very low: one inadequate study (results of study with no data provided are ignored)
IWADL/ APPADL – *(Physical) activities of daily living*[41, 42]	b. 12 out of 34 hypotheses confirmed; n=349	b. Inconsistent	b. High: one very good study
QOLID – *Physical endurance*[43]	a. 3 out of 6 hypotheses confirmed; n=30 b. 0 out of 3 hypotheses confirmed; n=210	a. Inconsistent b. Insufficient	a. Very low: one adequate study, sample size 30 b. High: one very good study
DQLCTQ – *Physical function*[44]	b. 1 out of 2 hypotheses confirmed; n=942	b. Inconsistent	b. High: one very good study
Diabetes-39 – *Energy and mobility (pilot version - 14 items)*[45]	b. No data provided	b. Indeterminate	b. No level of evidence: rating is indeterminate

Diabetes-39 – *Energy and mobility (15 items)*[45-51]	a. 8 out of 15 hypotheses confirmed; n=795 b. 11 out of 22 hypotheses confirmed; n=566	a. Inconsistent b. Inconsistent	a. High: two adequate studies (results of study with no data provided are ignored) b. High: two very good studies (results of study with no data provided are ignored)
Diabetes-39 Thai – *Energy and mobility (10 items)*[52]	a. 2 out of 5 hypotheses confirmed; n=397 b. 4 out of 6 hypotheses confirmed; n=397	a. Inconsistent b. Inconsistent	a. Moderate: one adequate study b. High: one very good study
Diabetes-39 German – *Physical impairment (7 items)*[53]	a. No data provided b. 4 out of 4 hypotheses confirmed; n=144	a. Indeterminate b. Sufficient	a. No level of evidence: rating is indeterminate b. High: one very good study
Diabetes-39 SF – *Energy and mobility (5 items)*[51]			
C-CWIS – *Physical symptoms and everyday living*[54]	a. No data provided b. 1 out of 1 hypotheses confirmed; n=131	a. Indeterminate b. Sufficient	a. No level of evidence: rating is indeterminate b. Very low: one inadequate study

RESPONSIVENESS **a=comparison to gold standard** **b=comparison with other instruments** **c=comparison between subgroups** **d=before and after intervention**			
PROM – *subscale*	**Summary or pooled result**	**Overall rating**	**Quality of evidence**
DFS – *Daily activities*[34]	c. 0 out of 2 hypotheses confirmed; n=264	c. Insufficient	c. Low: one doubtful study
DFS-SF – *Dependence/daily life*[35-39]	c. 2 out of 3 hypotheses confirmed; n=573	c. Inconsistent	c. High: three very good studies
PRO-DM-Thai – *Physical function*[40]			
IWADL/ APPADL – *(Physical) activities of daily living*[41, 42]	d. 2 out of 3 hypotheses confirmed; n=40	d. Inconsistent	d. Low: one very good study, sample size 40

QOLID – *Physical endurance*[43]			
DQLCTQ – *Physical function*[44]	c. 0 out of 2 hypotheses confirmed; n=328	c. Insufficient	c. High: one very good study
Diabetes-39 – *Energy and mobility (pilot version - 14 items)*[45]			
Diabetes-39 – *Energy and mobility (15 items)*[45-51]			
Diabetes-39 Thai – *Energy and mobility (10 items)*[52]			
Diabetes-39 German – *Physical impairment (7 items)*[53]	d. No data provided	d. Indeterminate	d. No level of evidence: rating is indeterminate
Diabetes-39 SF – *Energy and mobility (5 items)*[51]			
C-CWIS – *Physical symptoms and everyday living*[54]			

Appendix 12. Quality of PROM development and content validity

PROM	PROM design							Cognitive interview (CI) study[a]				Total PROM development	Content validity				
	General design requirements					Concept elicitation	Total PROM design	General design requirements	Comprehensibility	Comprehensiveness	Total CI study		Asking patients			Asking experts	
	Clear construct	Clear origin of construct	Clear target population for which the PROM was developed	Clear context of use	PROM developed in sample representing the target population			CI study performed in sample representting the target population					Relevance	Comprehensivness	Comprhensibility	Relevance	Comprehensiveness
DFS[34]	I	D	V	V	A	D	I	A	D		D	I					
DFS-SF[35, 37, 39]	V	D	V	V	A	D	D	A	D			D	I[b]		D[b]		
PRO-DM-Thai[40]	V	D	V	D	D	D	D	A	D		D	D					
IWADL/ APPADL[41]	V	D	V	V	V	D	D	V	?	?	D	D					
QOLID[43]	V	D	V	V	A	D	D	D	?	?	D	D					
DQLCTQ[44]	I	V	V	V	D	D	I					I					
Diabetes-39[45-48]	I	V	V	D	D	D	I					I			D[c]		
Diabetes-39 SF[51]	I	V	V	D	D	D	I					I					
C-CWIS[54]	I	D	V	V	I		I	A	D		D	I					

a Empty cells indicate that a CI study or content validity study (or part of it) was not performed, ? indicates that something was done but unclear what was done; b Content validity of the Chinese and Spanish version were assessed; c Content validity of the Arabic, Vietnamese and Portuguese version were assessed

Appendix 1. COSMIN definitions of domains, measurement properties and aspects of measurement properties[12]

Domain	Term: Measurement property	Term: Measurement property aspect	Definition
Reliability			The degree to which the measurement is free from measurement error
Reliability (extended definition)			The extent to which scores for patients who have not changed are the same for repeated measurement under several conditions: e.g., using different sets of items from the same OMI (internal consistency); over time (test-retest); by different persons on the same occasion (inter-rater); or by the same persons on different occasions (intra-rater)
	Internal consistency		The degree of interrelatedness among the items
	Reliability		The proportion of the total variance in the measurements which is due to 'true'† differences between patients
	Measurement error		The systematic and random error of a patient's score that is not attributed to true changes in the construct to be measured
Validity			The degree to which an OMI measures the construct(s) it purports to measure
	Content validity		The degree to which the content of an OMI is an adequate reflection of the construct to be measured
		Face validity	The degree to which (the items of) an OMI indeed seems to be an adequate reflection of the construct to be measured
	Construct validity		The degree to which the scores of an OMI are consistent with hypotheses (e.g., with regard to internal relationships, relationships to scores of other OMIs, or differences between relevant groups) based on the assumption that the OMI validly measures the construct to be measured
		Structural validity	The degree to which the scores of an OMI are an adequate reflection of the dimensionality of the construct to be measured
		Hypotheses testing	Idem construct validity
		Cross-cultural validity	The degree to which the performance of the items on a translated or culturally adapted OMI are an adequate reflection of the performance of the items of the original version of the OMI
	Criterion validity		The degree to which the scores of an OMI are an adequate reflection of a gold standard
Responsiveness			The ability of an OMI to detect change over time in the construct to be measured
	Responsiveness		Idem responsiveness
Interpretability*			The degree to which one can assign qualitative meaning (i.e., clinical or commonly understood connotations) to an OMI's quantitative scores or change in scores

COSMIN: COnsensus-based Standards for the selection of health Measurement INstruments

* Not considered a measurement property, but an important characteristic of a measurement instrument

† The word "true" must be seen in the context of the CTT, which states that any observation is composed of two components—a true score and error associated with the observation. "True" is the average score that would be obtained if the scale were given an infinite number of times. It refers only to the consistency of the score and not to its accuracy.[70]

Appendix 2. Search strategy

PUBMED search January 1, 2022

#1 Diabetes type 2

((Diabet*[tiab] AND (("non insulin"[tiab] AND depend*[tiab]) OR ("noninsulin"[tiab] AND depend*[tiab]) OR "type 2"[tiab] OR "type II" [tiab])) OR iddm[tiab] OR niddm[tiab] OR "glucose intolerance"[tiab] OR "insulin resistant"[tiab] OR "insulin resistance"[tiab])

#2 Modified filter for studies on measurement properties*

~~instrumentation[sh] OR methods[sh] OR~~ "Validation Studies"[pt] ~~OR "Comparative Study"[pt]~~ OR "psychometrics"[MeSH] OR psychometr*[tiab] OR clinimetr*[tw] OR clinometr*[tw] OR "outcome assessment (health care)"[MeSH] OR "outcome assessment"[tiab] OR "outcome measure*"[tw] OR "observer variation"[MeSH] OR "observer variation"[tiab] ~~OR "Health Status Indicators"[Mesh]~~ OR "reproducibility of results"[MeSH] OR reproducib*[tiab] OR "discriminant analysis"[MeSH] OR reliab*[tiab] OR unreliab*[tiab] OR valid*[tiab] OR "coefficient of variation"[tiab] ~~OR coefficient[tiab]~~ OR homogeneity[tiab] OR homogeneous[tiab] OR "internal consistency"[tiab] OR (cronbach*[tiab] AND (alpha[tiab] OR alphas[tiab])) OR (item[tiab] AND (correlation*[tiab] OR selection*[tiab] OR reduction*[tiab])) OR agreement[tw] OR precision[tw] OR imprecision[tw] OR "precise values"[tw] OR test-retest[tiab] OR (test[tiab] AND retest[tiab]) OR (reliab*[tiab] AND (test[tiab] OR retest[tiab])) OR stability[tiab] OR interrater[tiab] OR inter-rater[tiab] OR intrarater[tiab] OR intra-rater[tiab] OR intertester[tiab] OR inter-tester[tiab] OR intratester[tiab] OR intra-tester[tiab] OR interobserver[tiab] OR inter-observer[tiab] OR intraobserver[tiab] OR intra-observer[tiab] OR intertechnician[tiab] OR inter-technician[tiab] OR intratechnician[tiab] OR intra-technician[tiab] OR interexaminer[tiab] OR inter-examiner[tiab] OR intraexaminer[tiab] OR intra-examiner[tiab] OR interassay[tiab] OR inter-assay[tiab] OR intraassay[tiab] OR intra-assay[tiab] OR interindividual[tiab] OR inter-individual[tiab] OR intraindividual[tiab] OR intra-individual[tiab] OR interparticipant[tiab] OR inter-participant[tiab] OR intraparticipant[tiab] OR intra-participant[tiab] OR kappa[tiab] OR kappa's[tiab] OR kappas[tiab] OR repeatab*[tw] OR ((replicab*[tw] OR repeated[tw]) AND (measure[tw] OR measures[tw] OR findings[tw] ~~OR result[tw] OR results[tw]~~ OR test[tw] OR tests[tw])) OR generaliza*[tiab] OR generalisa*[tiab] OR concordance[tiab] OR (intraclass[tiab] AND correlation*[tiab]) OR discriminative[tiab] OR "known group"[tiab] OR "factor analysis"[tiab] OR "factor analyses"[tiab] OR "factor structure"[tiab] OR "factor structures"[tiab] ~~OR dimension*[tiab]~~ OR subscale*[tiab] OR (multitrait[tiab] AND scaling[tiab] AND (analysis[tiab] OR analyses[tiab])) OR "item discriminant"[tiab] OR "interscale correlation*"[tiab] OR error[tiab] OR errors[tiab] OR "individual variability"[tiab] OR "interval variability"[tiab] OR "rate variability"[tiab] OR (variability[tiab] AND (analysis[tiab] OR values[tiab])) ~~OR (uncertainty[tiab] AND (measurement[tiab] OR measuring[tiab]))~~ OR "standard error of measurement"[tiab] ~~OR sensitiv*[tiab]~~ OR responsive*[tiab] OR (limit[tiab] AND detection[tiab]) OR "minimal detectable concentration"[tiab] OR interpretab*[tiab] OR ((minimal[tiab] OR minimally[tiab] OR clinical[tiab] OR clinically[tiab]) AND (important[tiab] ~~OR significant[tiab]~~ OR detectable[tiab]) AND (change[tiab] OR difference[tiab])) OR (small*[tiab] AND (real[tiab] OR detectable[tiab]) AND (change[tiab] OR difference[tiab])) OR "meaningful change"[tiab] OR "ceiling effect"[tiab] OR "floor effect"[tiab] OR "Item response model"[tiab] OR IRT[tiab] OR Rasch[tiab] OR "Differential item functioning"[tiab] OR DIF[tiab] OR "computer adaptive testing"[tiab] OR "item bank"[tiab] OR "cross-cultural equivalence"[tiab]

#3 PROM filter (developed by the University of Oxford, see www.comin.nl)

(HR-PRO[tiab] OR HRPRO[tiab] OR HRQL[tiab] OR HRQoL[tiab] OR QL[tiab] OR QoL[tiab] OR quality of life[tw] OR life quality[tw] OR health index*[tiab] OR health indices[tiab] OR health profile*[tiab] OR health status[tw] OR ((patient[tiab] OR self[tiab] OR child[tiab] OR parent[tiab] OR carer[tiab] OR proxy[tiab]) AND ((report[tiab] OR reported[tiab] OR reporting[tiab]) OR (rated[tiab] OR rating[tiab] OR ratings[tiab]) OR based[tiab] OR (assessed[tiab] OR assessment[tiab] OR assessments[tiab]))) OR ((disability[tiab] OR function[tiab] OR functional[tiab] OR functions[tiab] OR subjective[tiab] OR utility[tiab] OR utilities[tiab] OR wellbeing[tiab] OR well being[tiab]) AND (index[tiab] OR indices[tiab] OR instrument[tiab] OR instruments[tiab] OR measure[tiab] OR measures[tiab] OR questionnaire[tiab] OR questionnaires[tiab] OR profile[tiab] OR profiles[tiab] OR scale[tiab] OR scales[tiab] OR score[tiab] OR scores[tiab] OR status[tiab] OR survey[tiab] OR surveys[tiab])))

(#1 AND #2 AND #3) NOT ("addresses"[Publication Type] OR "biography"[Publication Type] OR "case reports"[Publication Type] OR "comment"[Publication Type] OR "directory"[Publication Type] OR "editorial"[Publication Type] OR "festschrift"[Publication Type] OR "interview"[Publication Type] OR "lectures"[Publication Type] OR "legal cases"[Publication Type] OR "legislation"[Publication Type] OR "letter"[Publication Type] OR "news"[Publication Type] OR "newspaper article"[Publication Type] OR "patient education handout"[Publication Type] OR "popular works"[Publication Type] OR "congresses"[Publication Type] OR "consensus development conference"[Publication Type] OR "consensus development conference, nih"[Publication Type] OR "practice guideline"[Publication Type]) NOT ("animals"[MeSH Terms] NOT "humans"[MeSH Terms])

EMBASE search January 1, 2022

#1 diabetes type 2

(Diabet*:ti,ab AND (('non insulin':ti,ab AND depend*:ti,ab) OR (noninsulin:ti,ab AND depend*:ti,ab) OR 'type 2':ti,ab OR 'type II':ti,ab)) OR iddm:ti,ab OR niddm:ti,ab OR 'glucose intolerance':ti,ab OR 'insulin resistant':ti,ab OR 'insulin resistance':ti,ab

#2 Modified filter for studies on measurement properties*

~~'intermethod comparison'/exp OR~~ 'data collection method'/exp OR 'validation study'/exp OR 'feasibility study'/exp OR 'pilot study'/exp OR 'psychometry'/exp OR 'reproducibility'/exp OR reproducib*:ab,ti OR 'audit':ab,ti OR psychometr*:ab,ti OR clinimetr*:ab,ti OR clinometr*:ab,ti OR 'observer variation'/exp OR 'observer variation':ab,ti OR 'discriminant analysis'/exp OR 'validity'/exp OR reliab*:ab,ti OR valid*:ab,ti ~~OR 'coefficient':ab,ti~~ OR 'internal consistency':ab,ti OR (cronbach*:ab,ti AND ('alpha':ab,ti OR 'alphas':ab,ti)) OR 'item correlation':ab,ti OR 'item correlations':ab,ti OR 'item selection':ab,ti OR 'item selections':ab,ti OR 'item reduction':ab,ti OR 'item reductions':ab,ti OR 'agreement':ab,ti OR 'precision':ab,ti OR 'imprecision':ab,ti OR 'precise values':ab,ti OR 'test-retest':ab,ti OR ('test':ab,ti AND 'retest':ab,ti) OR (reliab*:ab,ti AND ('test':ab,ti OR 'retest':ab,ti)) OR 'stability':ab,ti OR 'interrater':ab,ti OR 'inter-rater':ab,ti OR 'intrarater':ab,ti OR 'intra-rater':ab,ti OR 'intertester':ab,ti OR 'inter-tester':ab,ti OR 'intratester':ab,ti OR 'intra-tester':ab,ti OR 'interobeserver':ab,ti OR 'inter-observer':ab,ti OR 'intraobserver':ab,ti OR 'intra-observer':ab,ti OR 'intertechnician':ab,ti OR 'inter-technician':ab,ti OR 'intratechnician':ab,ti OR 'intra-technician':ab,ti OR 'interexaminer':ab,ti OR 'inter-examiner':ab,ti OR 'intraexaminer':ab,ti OR 'intra-examiner':ab,ti OR 'interassay':ab,ti OR 'inter-assay':ab,ti OR 'intraassay':ab,ti OR 'intra-assay':ab,ti OR 'interindividual':ab,ti OR 'inter-individual':ab,ti OR 'intraindividual':ab,ti OR 'intra-individual':ab,ti OR 'interparticipant':ab,ti OR 'inter-participant':ab,ti OR 'intraparticipant':ab,ti OR 'intra-participant':ab,ti OR 'kappa':ab,ti OR 'kappas':ab,ti OR 'coefficient of variation':ab,ti OR repeatab*:ab,ti OR (replicab*:ab,ti OR 'repeated':ab,ti AND ('measure':ab,ti OR 'measures':ab,ti OR 'findings':ab,ti ~~OR 'result':ab,ti OR 'results':ab,ti~~ OR 'test':ab,ti OR 'tests':ab,ti)) OR generaliza*:ab,ti OR generalisa*:ab,ti OR 'concordance':ab,ti OR ('intraclass':ab,ti AND correlation*:ab,ti) OR 'discriminative':ab,ti OR 'known group':ab,ti OR 'factor analysis':ab,ti OR 'factor analyses':ab,ti OR 'factor structure':ab,ti OR 'factor structures':ab,ti OR 'dimensionality':ab,ti OR subscale*:ab,ti OR 'multitrait scaling analysis':ab,ti OR 'multitrait scaling analyses':ab,ti OR 'item discriminant':ab,ti OR 'interscale correlation':ab,ti OR 'interscale correlations':ab,ti OR ('error':ab,ti OR 'errors':ab,ti AND (measure*:ab,ti OR correlat*:ab,ti OR evaluat*:ab,ti OR 'accuracy':ab,ti OR 'accurate':ab,ti OR 'precision':ab,ti OR 'mean':ab,ti)) OR 'individual variability':ab,ti OR 'interval variability':ab,ti OR 'rate variability':ab,ti OR 'variability analysis':ab,ti OR 'standard error of measurement':ab,ti ~~OR sensitiv*:ab,ti~~ OR responsive*:ab,ti OR ('limit':ab,ti AND 'detection':ab,ti) OR 'minimal detectable concentration':ab,ti OR interpretab*:ab,ti OR (small*:ab,ti AND ('real':ab,ti OR 'detectable':ab,ti) AND ('change':ab,ti OR 'difference':ab,ti)) OR 'meaningful change':ab,ti OR 'minimal important change':ab,ti OR 'minimal important difference':ab,ti OR 'minimally important change':ab,ti OR 'minimally important difference':ab,ti OR 'minimal detectable change':ab,ti OR 'minimal detectable difference':ab,ti OR 'minimally detectable change':ab,ti OR 'minimally detectable difference':ab,ti OR 'minimal real change':ab,ti OR 'minimal real difference':ab,ti OR 'minimally real change':ab,ti OR 'minimally real difference':ab,ti OR 'ceiling effect':ab,ti OR 'floor effect':ab,ti OR 'item response model':ab,ti OR 'irt':ab,ti OR 'rasch':ab,ti OR 'differential item functioning':ab,ti OR 'dif':ab,ti OR 'computer adaptive testing':ab,ti OR 'item bank':ab,ti OR 'cross-cultural equivalence':ab,ti

#3 PROM filter (developed by the University of Oxford, see www.comin.nl)

(HR-PRO:ti,ab OR HRPRO:ti,ab OR HRQL:ti,ab OR HRQoL:ti,ab OR QL:ti,ab OR QoL:ti,ab OR 'quality of life':ti,ab OR 'life quality':ti,ab OR 'health index*':ti,ab OR 'health indices':ti,ab OR 'health profile*':ti,ab OR 'health status':ti,ab OR ((patient:ti,ab OR self:ti,ab OR child:ti,ab OR parent:ti,ab OR carer:ti,ab OR proxy:ti,ab) AND ((report:ti,ab OR reported:ti,ab OR reporting:ti,ab) OR (rated:ti,ab OR rating:ti,ab OR ratings:ti,ab) OR based:ti,ab OR (assessed:ti,ab OR assessment:ti,ab OR assessments:ti,ab))) OR ((disability:ti,ab OR function:ti,ab OR functional:ti,ab OR functions:ti,ab OR subjective:ti,ab OR utility:ti,ab OR utilities:ti,ab OR wellbeing:ti,ab OR 'well being':ti,ab) AND (index:ti,ab OR indices:ti,ab OR instrument:ti,ab OR instruments:ti,ab OR measure:ti,ab OR measures:ti,ab OR questionnaire:ti,ab OR questionnaires:ti,ab OR profile:ti,ab OR profiles:ti,ab OR scale:ti,ab OR scales:ti,ab OR score:ti,ab OR scores:ti,ab OR status:ti,ab OR survey:ti,ab OR surveys:ti,ab)))

#4 publicatie types

#3 AND ('article'/it OR 'article in press'/it OR 'review'/it)

#5 not animals

#4 NOT ([animals]/lim NOT [humans]/lim)

* Modified from Terwee et al.[28] The crossed-out search terms were left out because these terms, in combination with the search terms for diabetes, yielded too many abstracts to read.

Appendix 3. Criteria for good measurement properties

Measurement property	Rating	Criteria
Structural validity*	+	**CTT:** EFA/PCA: factor loadings of each item on its factor is at least 0.30 *AND* maximum 10% of the items load on more than one factor *AND* minimum explained variance is 50% and structure is in line with the theory about the construct to be measured *OR* results on scree plot or Kaiser criterion (Eigenvalues >1) are in line with the theory about the construct to be measured CFA: CFI or TLI or comparable measure >0.95 *OR* RMSEA <0.06 *OR* SRMR <0.08 **IRT/Rasch:** no violation of <u>unidimensionality:</u> CFI or TLI or comparable measure >0.95 *OR* RMSEA <0.06 *OR* SRMR <0.08 *AND* no violation of <u>local independence:</u> residual correlations among the items after controlling for dominant factor <0.20 *OR* Q3's <0.37 *AND* no violation of <u>monotonicity:</u> adequate looking graphs *OR* item scalability >0.30 *AND* adequate <u>model fit:</u> IRT: χ^2 >0.01 Rasch: infit and outfit mean squares ≥0.5 and ≤1.5 *OR* Z-standardized values >-2 and <2
	?	**CTT:** not all information for '+' reported **IRT/Rasch:** model fit not reported
	-	Criteria for '+' not met
Internal consistency	+	At least low evidence for sufficient structural validity *AND* Cronbach's alpha(s) ≥0.70 for each unidimensional scale or subscale
	?	Criteria for "at least low evidence for sufficient structural validity" not met
	-	At least low evidence for sufficient structural validity *AND* Cronbach's alpha(s) <0.70 for each unidimensional scale or subscale
Reliability	+	ICC or (weighted) kappa or Pearson/Spearman correlation ≥0.70
	?	ICC or (weighted) kappa or Pearson/Spearman correlation not reported
	-	ICC or (weighted) kappa or Pearson/Spearman correlation <0.70
Measurement error	+	SDC or LoA <MIC
	?	MIC not defined
	-	SDC or LoA > MIC
Hypotheses testing for construct validity	+	≥75% of the results is in accordance with predefined hypotheses
	?	No hypotheses defined (by the review team)
	-	≥75% of the results is not in accordance with predefined hypotheses
Cross-cultural validity\ measurement invariance	+	No important differences found between group factors (such as age, gender, language) in multiple group factor analysis *OR* no important DIF for group factors (McFadden's R^2 <0.02)
	?	No multiple group factor analysis *OR* DIF analysis performed
	-	Important differences between group factors *OR* DIF was found
Criterion validity	+	Correlation with gold standard ≥0.70 *OR* AUC ≥0.70
	?	Not all information for '+' reported
	-	Correlation with gold standard <0.70 *OR* AUC <0.70
Responsiveness	+	≥75% of the results is in accordance with predefined hypotheses *OR* AUC ≥0.70

	?	No hypotheses defined (by the review team)
	-	≥75% of the results is not in accordance with predefined hypotheses *OR* AUC <0.70

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

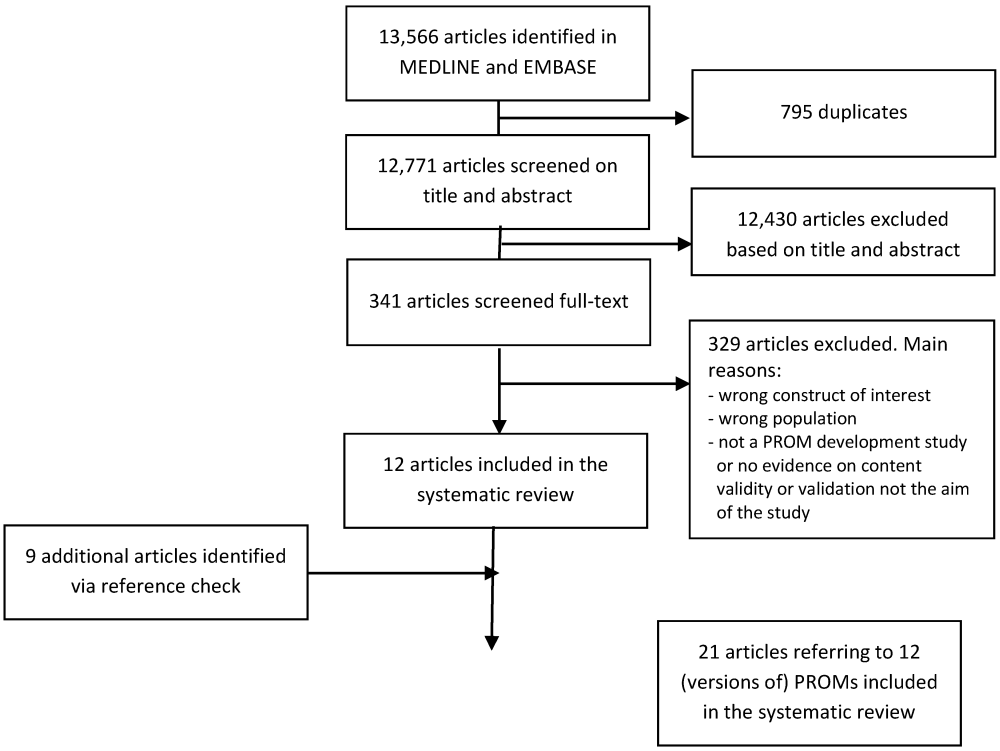
*Standard 1 in Box 3 in the COSMIN Risk of Bias checklist[32] was rated very good if CFA was performed, adequate if EFA was performed, doubtful if PCA was performed and inadequate if none of the previous was performed.

Appendix 4. Approach for grading the quality of the evidence

Grade factor	Downgrading	Definition
Risk of bias	0	Multiple studies of at least adequate quality *OR* one study of very good quality
	-1	Only one study of adequate quality *OR* multiple studies of doubtful quality
	-2	Only one study of doubtful quality *OR* multiple studies of inadequate quality
	-3	Only one study of inadequate quality
Imprecision (not for content validity, structural validity, and cross-cultural validity\ measurement invariance)	0	Total sample size of all studies >100
	-1	Total sample size of all studies 50-100
	-2	Total sample size of all studies <50
Inconsistency	0	Results are consistent *OR* results are summarized and rated per subset of studies, and subsequently graded
	-1	Overall rating based on the majority of consistent results
Indirectness	0	Does not occur; definitions for construct and/or target population have been stated in the inclusion criteria

0: high, -1: moderate, -2: low, -3: very low; Per protocol of the COSMIN guideline for systematic reviews: the quality of evidence for internal consistency cannot be higher than the quality of evidence for structural validity[23]

Appendix 5. Flowchart of the results from the search strategy



Appendix 6. Study populations involved in PROM development and content validity studies (marked with *)

PROM	Language	Country	Patients	Patient input	Mean age [range] yr	Gender % female	Disease characteristics	Professionals	Characteristics of professionals	Professional input
DFS[42]	English	UK	10 + 14 + 12	Concept elicitation and pilot testing relevance and comprehensibility	61 [46-74]	33-40%	Diabetic foot ulcers	0		
DFS-SF[34]	English	US	0					0		
DFS-SF[38]*	Chinese	China	6	Comprehensibility and relevance				0		
DFS-SF[61]*	Spanish	Spain	0					?		Evaluation content validity
PRO-DM-Thai[35]	Thai	Thailand	12 + 15	Concept elicitation/ comprehensibility, face validity	61 [49-70]	50%	T2DM, disease duration 11 [2-20] yr	9 + 17	3 physicians, 2 nurses, 2 pharmacists, and 2 nutritionists	Concept elicitation / relevance, comprehensiveness, comprehensibility
IWADL/ APPADL[36]	English	US	54 + 24	Concept elicitation, comprehensibility			T2DM and BMI of 25–40 kg/m2	0		
QOLID[37]	English/ Hindi[a]	India	20	Concept elicitation, comprehensibility			T2DM	8	4 clinicians and 4 diabetes educators	Relevance, comprehensibility
DQLCTQ[44]	English	US	30	Rating domains			T1DM (n=23), T2DM (n=7)	11	Clinicians / experts on HRQL research	Rating domains / evaluating face- and content validity of draft PROM
Diabetes-39[43]	English	US	?	Concept elicitation			Diabetes	?	Physicians, certified diabetes educators, pharmacists	Concept elicitation
Diabetes-39[39]*	Arabic	Jordan	30	Comprehensibility				0		
Diabetes-39[40]*	Vietnamese	Vietnam	10	Comprehensibility				0		
Diabetes-39[41]*	Portuguese	Brazil	4	Comprehensibility				0		

Diabetes-39 SF[48]	Chinese	Taiwan	0			0		
C-CWIS[47]	Chinese	China	20	Pilot testing	Diabetic foot ulcers	5	2 majored in medicine, 3 majored in nursing	Transcultural adjustment

a Language unsure

Appendix 7. Characteristics of included study populations for the assessment of measurement properties

	Population			**Disease characteristics**		**Instrument administration**			
PROM	**N**	**Mean age (SD) [range] yr**	**Gender % female**	**Disease**	**Disease duration mean (SD) yr**	**Setting**	**Country**	**Language**	**Response rate**
DFS[42]	48 + 54 + 71 = 173 Internal consistency: 48 Structural validity: 326 Known-groups: 102 (48+54) Responsiveness: 264	58 (11), 65 (12), 55 (15) N=326: ? N=264: ?	27%, 28%, 34% N=326: ? N=264: ?	T2DM and T1DM with neuropathic or mixed neuropathic/ ischemic foot ulcer	17 (11), 18 (12), 12 (9) N=326: ? N=264: ?	Diabetic foot centres	UK N=326: UK, US, Belgium, Denmark, France, Italy, Netherlands	English, and other	
DFS-SF study 1[34]	180	[27-87]	26%	Chronic, neuropathic, full-thickness, diabetic foot ulcers		Clinical trials	Belgium (16), Denmark (11), France (3), Germany (5), Italy (3), Netherlands (5), UK (243), US (40)	English, and other	
DFS-SF study 2[34]	252 + 95 = 347 Responsiveness: 252	[29-88], [33-83]	30%, 26%	Chronic, neuropathic, full-thickness, diabetic foot ulcers		Clinical trials	Austria (23), France (62), Germany (67), Greece (34), Italy (18), Netherlands (66), Switzerland (15) UK (16), US (95)	English, and other	
DFS-SF Polish[45]	212	63 (10)	28%	Diabetes (T2DM 86%)	18 (12)	Ambulatory setting of university	Poland	Polish	

						hospital			
DFS-SF Chinese[38]	60	71 (11)	40%	Current/healed diabetic foot ulcers	17 (10)	Outpatient/inpatient hospital	Hong Kong	Hong-Kong Chinese	
DFS-SF Greek[60]	110	60 (12)	39%	Diabetes (T2DM 86%) with diagnosed diabetic foot ulcer	17 (7)	General hospital	Greece	Greek	88%
DFS-SF Spanish[61]	141	68 (13)	33%	Diabetes (T2DM 95%) with new-onset diabetic foot ulcer		Diabetic foot unit	Spain	Spanish	
PRO-DM-Thai[35]	500 Construct validity: 200	66 (11) [32-90] N=200: 67 (11) [32-89]	67% N=200: 68%	T2DM (well-controlled and uncontrolled)	15 (8) N=200: 16 (8)	Outpatient clinic university hospital	Thailand	Thai	
IWADL/APPADL[36]	349	59 (9)	56%	T2DM with BMI 30-40 kg/m2		Internet panel	USA	English	
IWADL/APPADL[46]	106 Responsiveness: 40	52 (10) N=40: 52 (12)	69% N=40: 65%	T2DM with BMI at least 30 kg/m2		Weight loss centres, university-based weight loss program, or specialty clinic	USA	English	89%
QOLID[37]	150 Comparison instruments: 30 Known-groups: 210[a]	54 (7) [35-65] N=30 and N=210: similar to N=150	40% N=30 and N=210: similar to N=150	T2DM	11 (6) N=30 and N=210: similar to N=150	Diabetes centre	India	English/Hindi[b]	33% (150/460); 68% 30/44; 73% 30/41
DQLCTQ[44]	942 Internal	Type I diabetes: 34, type II	43% N=909: ?	T1DM (n=468) and T2DM (n=474)[c]	13 N=909: ?	Clinical trials	Canada (72), France (84), Germany (188),	English, German, French	

	consistency: 909 Reliability: 58 Responsiveness: 328	diabetes: 58 N=909: ? N=58: ? N=328: ?	N=58: ? N=328: ?		N=58: ? N=328: ?		US (598		
Diabetes-39 study 1[43]	516	52 (16)	55%	T1DM (33%) and T2DM (68%)	14 (11)	Diabetes care centre	US	English	52%
Diabetes-39 study 2[43]	165 + 262 = 427	62 (18), 55 (13)	55%, 65%	T1DM (20%, 10%) and T2DM (81%, 90%)	12 (11), 10 (8)	General practice, hospital diabetes clinic	US	English	70%, 41%
Diabetes-39 Arabic[39]	368	57 (10)	56%	T2DM	9 (7)	Outpatient clinic of university hospital	Jordan	Arabic	92%
Diabetes-39 Vietnamese[40]	286	20% <50, 45% 60-65, 36% >65	64%	T2DM	6 (5)	Tertiary hospital	Vietnam	Vietnamese	
Diabetes-39 Portuguese[41]	52	63 (9) [45-84]	65%	T2DM	9 (4)	Basic health service	Brazil	Portuguese	
Diabetes-39 Spanish[51]	249	5% <40, 61% 40-59, 35% ≥60, Males: 52.5 [24-75] years), Females: 55.7 [34-91]	63%	T2DM	9 (8)	Family medicine unit	Mexico	Spanish	96%
Diabetes-39 Taiwan[59]	280	63 (11)	47%	T1DM (1%) and T2DM (99%)	9 (6)	Outpatient clinic of teaching hospital	Taiwan	Chinese	
Diabetes-39 Taiwan[48]	265[d]			T1DM and T2DM		Diabetes clinics of a teaching	Taiwan	Chinese	

						hospital			
Diabetes-39 Thai[49]	397	58 (11)	74%	Diabetic people	6 (6)	Community hospital	Thailand	Thai	93%
Diabetes-39 German[50]	144 Reliability: 72 Responsiveness: 62-66	57 (8) N=72: ? N=62-66: ?	52% N=72: ? N=62-66: ?	T2DM	13 (10) N=72: ? N=62-66: ?	Special hospital/outpatient clinic for people with diabetes	Germany	German	85%
Diabetes-39 SF[48]	265[d]			T1DM and T2DM		Diabetes clinics of a teaching hospital	Taiwan	Chinese	
C-CWIS[47]	131 Internal consistency: 20 & 131	68 (11) N=20: ?	34% N=20: ?	T2DM with diabetic foot ulcers	14 (8) N=20: ?	Outpatient/inpatient diabetic foot of an integrated hospital	China	Chinese	

a N=60 for HbA1C values; b Language unsure; c Analyses conducted separately for each group; pooled analyses performed because results were similar; d Sample from study of Huang[59]

Appendix 8. Information on feasibility of PROMs

PROM	Type and ease of administration	Length of instrument[a]	Response options included subscale	Completion time	Patient's required mental and physical ability level	Ease of score calculation	Copyright
DFS[42]	Self-report	11 subscales, 58 items: Leisure (5); Physical health (6); **Daily activities (6)**; Emotions (17); Noncompliance (2); Family (5); Friends (5); Treatment (4); Satisfaction (1); Positive attitude (5); Financial (2)	Daily activities: 1 = none of the time, 2 = a little bit of the time, 3 = some of the time, 4 = most of the time, and 5 = all of the time			Scores are based on the sum of items associated with a subscale if at least 50% of the items in a scale are completed. When necessary, raw item scores are reverse coded so that the minimum possible score (1) represents the worst quality of life, and the maximum possible score (5) represents the best quality of life (all items except in the positive attitude subscale). Each subscale is scored from 0 to 100, higher scores indicate better quality of life.	Johnson & Johnson Research & Development, LCC
DFS-SF[34,38,45,60,61]	Self-report/interview-based	6 subscales, 29 items: Leisure (5); **dependence/ daily life (5)**; negative emotions (6); physical health (5); worried about ulcers/feet (4); bothered by ulcer care (4)	Dependence/ daily life: 1 = none of the time, 2 = a little of the time, 3 = some of the time, 4 = most of the time, and 5 = all of the time	12.5 minutes for interview-based administration		Scores are based on the sum of items associated with a subscale if at least 50% of the items in a scale are completed. In case of item-level missing data <50%, the subscale score is calculated by substituting the mean item score for the missing item values. Raw item scores are reverse coded so that the	Johnson & Johnson Research & Development, LCC

						minimum possible score (1) represents the worst quality of life, and the maximum possible score (5) represents the best quality of life. Each subscale is scored from 0 to 100, higher scores indicate better quality of life.	
PRO-DM-Thai[35]	Self-report/interview-based	7 subscales, 44 items: **Physical function (5);** symptoms (7); Psychological wellbeing (5); Self-care management (12); Social wellbeing (5); Global judgements of health (5); Satisfaction with care and flexibility of treatment (5)	Unknown, PROM could not be retrieved	30 minutes		Not reported.	
IWADL/APPADL[36,46]	Self-report	1 subscale, 7 items: **Physical activities of daily living (7)**	1-5: 1 = unable to do, 5 = not at all difficult	<5 minutes	Flesch Kincaid reading level: 9th grade	Total scores are derived by adding item scores (minimum = 1, maximum = 5) and then dividing by the number of items, so that the minimum and maximum total scores are 1 and 5, respectively. Total scores can be transformed to 0-100. Higher scores correspond to greater ability to do physical daily activities.	Publicly available
QOLID[37]	Interview-based	8 subscales, 34 items: Role limitations due to physical health (social life, work, traveling) (6); **Physical endurance (6);**	1-5: 1 referring to poorest outcome, 5 to best outcome	Mean: 7.8 minutes, SD: 2.8 minutes		A score for each domain is calculated by adding items' scores after mean imputation for 'not applicable' values. Each	

		General health (3); Treatment satisfaction (4); Symptom botherness (3); Financial worries (4); Emotional/ mental health (5); Diet advise tolerance (3)			domain score is standardized by dividing by the maximum possible domain score and multiplying by 100. All domain scores are added and divided by 8 (the number of domains) to obtain an overall score. Standardized scores range 0-100.	
DQLCTQ[44]	Self-report	8 subscales, 57 items: **Physical function (6)**; Energy/fatigue (5); Health distress (6); Mental health (5); Satisfaction (DQOL – 18, excl. 3 skip pattern questions); Treatment satisfaction (3); Treatment flexibility (10); Frequency of symptoms (7)	Physical function: 1 = limited for more than four weeks, 2 = limited for four weeks or less, 3 = not limited at all	10 minutes	The average of a domain is computed by summing up scores within the domain and dividing the sum by the number of items in the domain. If 50% or more of the items are missing, the average score should not be calculated, and the domain score is treated as missing. Domain scores are converted to a 100-point scale. Higher scores indicate better quality of life.	Yes
Diabetes-39[39-41,43,48-51,59]	Self-report/interview-based	5 subscales, 39 items: **Energy and mobility (15)**; Diabetes control (12); Anxiety and worry (4); Social burden (5); Sexual functioning (3)[b]	English/Arabic: VAS marked 1-7: 1 = not affected at all, 7 = extremely affected Other: 1-7: 1 = not affected at all, 7 = extremely affected	10-15 minutes	English/Arabic: Respondents place an 'X' on a modified visual analogue scale ranging from 1 (= not affected at all) to 7 (= extremely affected). The response is measured to the nearest quarter of a centimeter. Any response falling between two of the quarter graduations is rounded to the higher quarter of a centimeter. If	

				more than 7 items are missing, the questionnaire is not analyzed. In case of 7 or less missing items, the modal response within each subscale served as a proxy for missing data. Each scale score is transformed to 0-100, with 0 indicating the least impact on quality of life and 100 indicating the most impact. Other: Respondents place an 'X' in one of the boxes numbered 1 to 7, which are on a horizontal bar. The number marked, without any 0.5 point approximation, for each subscale is summed, and then transported to a scale from 0 to 100, with a higher score indicating greater impact on quality of life.[c]
Diabetes-39 SF[48]	Interview-based	5 subscales, 22 items: **Energy and mobility (5)**; Diabetes control (5); Anxiety and worry (4); Social burden (5); Sexual functioning (3)	1-7: 1 = not affected at all, 7 = extremely affected	Scores are given on a 7-point Likert scale. Subscale scores are calculated by summing all responses, high scores represent poor quality of life.
C-CWIS[47]	Self-report/interview-based	3 subscales, 25 items: **Physical symptoms and everyday living (5)**; Social life (7); Well-being (6)	1-5: 1 = not at all, 5= always	Total item scores include patient's perception of the experience and the associated stress. To calculate scale scores, the item scores and number of

			sub-evaluations are summated for each scale (unclear how this is exactly done, validated formula is used), subscale scores range from 0-100, higher scores indicate better health-related quality of life.

a Bold subscales measure physical functioning; b Thai version: 6 subscales and 39 items: Energy and mobility (10), Diabetes control (13), Anxiety and worry (4), Social burden (6), Sexual functioning (3), Other health problems and diabetic complications (3); German version: 5 subscales, 39 items: Diabetes and treatment (7), Physical impairment (7), Social stress (5), Physical illness (5), Sexual problems (3), subscale unknown for remaining 12 items; c Vietnamese version: If more than 4 items are missing (except from the sexual functioning domain), the questionnaire is not analyzed. For the energy and mobility scale, of more than 3 items are missing, a scale score is not calculated. If 3 of less items are missing, the missing value is replaced by the mean scale score

Appendix 9. Information on interpretability of PROMs

PROM – *subscale*	Distribution of scores in the study population	Percentage of missing items or percentage of missing scores	Floor and ceiling effects	Scores and change scores available for relevant (sub)groups	Minimal important change (MIC) or minimal important difference (MID)
DFS – *Daily activities*[42]				Healed ulcer: ~69, Current ulcer: ~63[a]	
DFS-SF – *Dependence/ daily life*[34,38,45,60,61]	Ref 45: mean=47.7, median =50.0, SD=29.3 Ref 38: mean=71.4, median =85.0, SD=32.9 Ref 60: mean=56.3, median=55.0, SD=25.7	Ref 45: 0.0-1.5% Ref 38: 0%	Ref 45: 7.8% floor, 2.9% ceiling Ref 38: 5% floor, 30% ceiling Ref 60: 0.9% floor, 5.5% ceiling	Ref 34: Pre vs. post closure of target ulcer change score – study 1: +3.7; study 2 +10.0 Ref 38: Healed vs. unhealed ulcer Change score: +13.9 Ref 60: >1 complication: 44.7, 1 complication: 55.8, No complication: 69.5	
PRO-DM-Thai – *Physical function*[35]					
IWADL/APPADL – *(Physical) activities of daily living*[36,46]	Ref 36: Mean=3.3, SD=1.1 Ref 46: Mean=3.3, SD=1.0		Ref 36: 4-31% floor per item, 8-32% ceiling per item Ref 46: 6% floor effect, 11% ceiling effect		Ref 46: If transformed to 0-100: SEM = 6.3; MIC based on weight loss: 13.6; MIC based on ability to perform daily physical activities: 9.8
QOLID – *Physical endurance*[37]	Mean=25.3, SD=5.5 Standardized mean =84.3, SD=18.4	20.7%		HbA1c ≤8: 87.1, HbA1c >8: 81.8; Insulin: 81.6, Non-insulin: 86.1; Comorbidity ≤1: 77.9, Comorbidity >1: 88.6; Male: 89.4, Female: 76.4	
DQLCTQ – *Physical function*[44]				HbA1c tight: 89.3, HbA1c poor: 85.1; Type 1: 94.7, Type 2: 77.9; Male: 87.9, Female: 84.1; Good control: 88.2, Poor control: 81.3	
Diabetes-39 – *Energy and mobility (pilot version - 14*		Study 1 - Total questionnaire: 0.3-0.45%			

items)[43]				
Diabetes-39 – *Energy and mobility (15 items)*[39-41,43,51,59]	Ref 39: mean=50.5, SD=21.1 Ref 40: median=41.1, 25th percentile =22.2, 75th percentile =60.0, range=0.0-91.1 Ref 41: median=51.5, mean=48.8, SD=14.5, range (possible: 15-105) =20-83 Ref 51: median=30, 25th percentile =16, 75th percentile =50	Ref 43: Study 2 - Total questionnaire: 0.3-0.45%	Ref 39: 0.3% ceiling, 0.8% floor Ref 41: 1.9 % floor, 1.9% ceiling	Ref 40: Male: 40.0, Female: 43.3; Insulin: 46.1, Non-insulin: 40.0; Comorbidity: 43.3, No comorbidity: 26.1 Complication: 49.4, No complication: 37.8 Ref 51: Male: 27, Female: 35
Diabetes-39 Thai – *Energy and mobility (10 items)*[49]		Total questionnaire: <2%		Comorbidity: 28.8, No comorbidity: 26.0 Insulin: 35.8, Non-insulin: 26.6; Complications: 37.0, No complications: 27.1
Diabetes-39 German – *Physical impairment (7 items)*[50]		N=1-4 for all subscales		Insulin: 18.4, No insulin: 15.8; ≤1 Complication: 11.5, >1 Complication: 20.4
Diabetes-39 SF – *Energy and mobility (5 items)*[48]				
C-CWIS – *Physical symptoms and everyday living*[47]	Item score ranges: mean=4.98-8.78, SD=1.79-2.58			

a Read from histogram

Appendix 10. Extensive results of studies on measurement properties

PROM – *subscale*	**Country (language) in which the PROM was evaluated**	**Structural validity**			**Internal consistency**			**Cross-cultural validity\ measurement invariance**			**Reliability**		
		n	**Meth qual**	**Result (rating)**	**n**	**Meth qual**	**Result (rating)**	**n**	**Meth qual**	**Result (rating)**	**n**	**Meth qual**	**Result (rating)**
DFS – *Daily activities*[42]	English, and other	326	Inade quate	Daily life (3 items) (?) and dependence (4 items) (?)	48	Inade quate	Daily activities: α^{a}=0.85 (?)				?	Inadequa te	Daily life: r=0.68 (-) Dependence: r=0.84 (+)
Pooled or summary result (overall rating)		**326**		**Daily life (3 items) and dependence (4 items) (?)**	**48**		**Daily activities: α=0.85 (?)**				**?**	**Very low**	**Daily life: r=0.68 (-) Dependence: r=0.84 (+)**
DFS-SF study 1 – *Dependence /daily life*[34]	English, and other	180	Inade quate	Dependence/ daily life (5 items): factor loadings 0.53-0.85; one cross- loader in whole instrument (?)	180	Very good	Dependence /daily life: α=0.88 – 0.91[b] (+)				?	Inadequa te	Dependence /daily life: ICC=0.77 (+)
DFS-SF study 2 – *Dependence /daily life*[34]	English, and other	347	Very good	Dependence/ daily life (5 items): non-normed fit index = 0.93; CFI = 0.94; incremental fit index = 0.94; goodness of fit index = 0.87; RMSEA = 0.058 (+)	347	Very good	Dependence /daily life: α=0.85 - 0.88[c] (+)				?	Inadequa te	Dependence /daily life: ICC=0.74 (+)
DFS-SF Polish – *Dependence*	Polish				212	Very good	Dependence /daily life:	212	Inadeq uate	DIF in item 3c			

/daily life[45]							α=0.90 (+)			for sex, (p-value Chi2 = 0.02) 6 variables were tested in 5 items (+)			
DFS-SF Chinese – *Dependence /daily life*[38]	Hong Kong Chinese				60	Very good	Dependence /daily life: α=0.89 (+)						
DFS-SF Greek – *Dependence /daily life*[60]	Greek				110	Very good	Dependence /daily life: α=0.87 (+)						
DFS-SF Spanish – *Dependence /daily life*[61]	Spanish	141	Inade quate	Dependence/ daily life (5 items): CFA: CFI = 0.844; RMSEA = 0.095; SRMR = 0.093 (-) EFA: 65.5% total variance explained; dependence/ daily life scale: factor loadings 0.58-1.00 (+)	141	Very good	Dependence /daily life: α=0.87 (+)				141	Doubtful	Dependence /daily life: ICC cons=0.77 (+)
Pooled or summary result (overall rating)		**347**	**High**	**Dependence/ daily life (5 items) (+)**	**105 0**	**High**	**Dependence /daily life: α=0.85-0.91**	**212**	**Very low**	**DIF for sex in one item**	**>141**	**Low**	**Dependence /daily life: ICC=0.74-0.77 (+)**

							(+)			**(+)**			
PRO-DM-Thai – *Physical function*[35]	Thai	500	Very good	Physical function (5 items): Goodness-of-Fit index 0.998; Adjusted Goodness-of-Fit index 0.991; RMSEA: 0.000 (+)	500	Very good	Physical function: α=0.82 (+)						
Pooled or summary result (overall rating)		**500**	**High**	**Physical function (5 items) (+)**	**500**	**High**	**Physical function: α=0.82 (+)**						
IWADL/APPADL – *(Physical) activities of daily living*[36]	English	349	Doubtful	(physical) activities of daily living (7 items): 73% variance explained; factor loadings 0.82-0.90; eigenvalue 5.1 (+)	349	Very good	(physical) activities of daily living: α=0.94 (+)						
IWADL/APPADL – *(Physical) activities of daily living*[46]	English				106	Very good	(physical) activities of daily living: α≥0.89[d] (+)				106	Adequate	(physical) activities of daily living: ICC agr=0.91 (+)
Pooled or summary result (overall rating)		**349**	**Low**	**(physical) activities of daily living (7 items) (+)**	**455**	**Low[e]**	**(physical) activities of daily living: α≥0.89 (+)**				**106**	**Moderate**	**(physical) activities of daily living: ICC agr=0.91 (+)**
QOLID – *Physical*	English/Hindi[f]	150	Inadequate	Physical endurance (6	150	Very good	Physical endurance:						

endurance[37]				items): 5.9%variance explained (total 49.9%); factor loadings 0.52 – 0.72 (+)			α=0.85 (?)						
Pooled or summary result (overall rating)		**150**	**Very low**	**Physical endurance (6 items): (+)**	**150**		**Physical endurance: α=0.85 (?)**						
DQLCTQ – *Physical function*[44]	English, German, French				909	Doubtful	Physical function: α=0.85 (?)				58	Adequate	Physical function: ICC=0.83 (+)
Pooled or summary result (overall rating)					**909**		**Physical function: α=0.85 (?)**				**58**	**Low**	**Physical function: ICC=0.83 (+)**
Diabetes-39 – *Energy and mobility (pilot version - 14 items)*[43]	English	516	Adequate	Energy and mobility (14 items): Seven factors with eigenvalue >1; 77% total variance explained; item loadings >0.50 (+)	516	Very good	All subscales: α=0.81-0.92 (+)						
Pooled or summary result (overall rating)		**516**	**Moderate**	**Energy and mobility (14 items) (+)**	**516**	**Moderate**[e]	**All subscales: α=0.81-0.92 (+)**						
Diabetes-39 – *Energy and mobility (15 items)*[43]	English	427	Adequate	Energy & mobility (15 items): Five factors with eigenvalue >1;	427	Very good	Energy and mobility: α=0.93 (+)						

				90% total variance explained; items loadings >0.50 (+)									
Diabetes-39 Arabic – *Energy and mobility (15 items)*[39]	Arabic	368	Doubtful	Energy & mobility (15 items): 83% total variance explained; item-scale correlations >0.40 (+)	368	Very good	Energy and mobility: α=0.89 (+)						
Diabetes-39 Vietnamese – *Energy and mobility (15 items)*[40]	Vietnamese				286	Very good	Energy and mobility: α=0.92 (+)				286[g]	Doubtful	Energy and mobility: ICC=0.91 (+)
Diabetes-39 Portuguese – *Energy and mobility (15 items)*[41]	Portuguese				52	Very good	Energy and mobility: α=0.79 (+)						
Diabetes-39 Spanish – *Energy and mobility (15 items)*[51]	Spanish				249	Very good	Energy and mobility: α=0.92 (+)				249	Doubtful	Energy and mobility: spearman r=0.84 (+)
Diabetes-39 Taiwan – *Energy and mobility (15 items)*[59]	Chinese				280	Very good	All subscales: α=0.82-0.93 (+)						
Diabetes-39 Taiwan –	Chinese	265	Adequate	Energy and mobility (15									

Energy and mobility (15 items)[48]				items): CFI: 0.92; TLI: 0.98; RMSEA: 0.085 (+)									
Pooled or summary result (overall rating)		**1060**	**High**	**Energy and mobility (15 items) (+)**	**1662**	**High**	**Energy and mobility: α=0.79-0.93 (+)**				**535**	**Moderate**	**Energy and mobility: spearman r=0.84, ICC=0.91 (+)**
Diabetes-39 Thai – *Energy and mobility (10 items)*[49]	Thai	397	Adequate	Energy and mobility (10 items): Factor loadings 0.28-0.80; 1 cross-loader in scale; total variance explained 62.1% (-)	397	Very good	Energy and mobility: α=0.94 (?)						
Pooled or summary result (overall rating)		**397**	**Moderate**	**Energy and mobility (10 items) (-)**	**397**		**Energy and mobility: α=0.94 (?)**						
Diabetes-39 German – *Physical impairment (7 items)*[50]	German	144	Inadequate	Physical impairment (7 items): Factor loadings >0.50 (?)	144	Doubtful	Physical impairment: α=0.84 (?)				72[6]	Doubtful	Physical impairment: 0.88[h] (?)
Pooled or summary result (overall rating)		**144**		**Physical impairment (7 items) (?)**	**144**		**Physical impairment: α=0.84 (?)**				**72**		**Physical impairment: 0.88 (?)**
Diabetes-39 SF – *Energy and mobility (5 items)*[48]	Chinese	265	Very good	Energy and mobility (5 items): CFI: 0.966; TLI: 0.993; RMSEA: 0.058 (+)									
Pooled or summary result		**265**	**High**	**Energy and**									

(overall rating)				**mobility (5 items) (+)**									
C-CWIS – *Physical symptoms and everyday living*[47]	Chinese	131	Adequate	Physical symptoms and everyday living (12 items): Factor loadings 0.53-0.78 (one item with loading 0.16 loaded on a different factor but was retained in its original factor); 39.4% variance explained (total 57.2%); eigen value 9.9 (+)	Pilot: 20 Formal: 131	Very good	Physical symptoms and everyday living: Pilot: α=0.73 (+) Formal: α=0.92 (+)						
Pooled or summary result (overall rating)		**131**	**Moderate**	**Physical symptoms and everyday living (12 items): (+)**	**151**	**Moderate[e]**	**Physical symptoms and everyday living: α=0.73-0.92 (+)**						

Appendix 10. Continued

PROM – *subscale*	**Country (language) in which the PROM was evaluated**	**Measurement error**	**Criterion validity**	**Hypotheses testing for construct validity a=comparison with other instruments b=comparison between subgroups**	**Responsiveness a=comparison to gold standard b=comparison with other instruments c=comparison between subgroups d=before and after intervention**

		n	Meth qual	Result (rating)	n	Meth qual	Result (rating)	n	Meth qual	Result (rating)	n	Meth qual	Result (rating)
DFS – *Daily activities*[42]	English, and other							a. 173 b. 102[i]	a. Adequate b. Very good	a. Results in line with 4 hypos (4+); results not in line with 1 hypo (1-) b. No data provided (?)	c. 264	c. Doubtful	c. Results not in line with 2 hypos (2-)
Pooled or summary result (overall rating)								**a. 173 b. 102**	**a. Moderate**	**a. 4+ and 1- (+) b. (?)**	**c. 264**	**c. Low**	**c. 2- (-)**
DFS-SF study 1 – *Dependence /daily life*[34]	English, and other							a. 180	a. Adequate	a. Results in line with 2 hypos (2+); results not in line with 3 hypos (3-)	c. 180	c. Very good	c. Results not in line with 1 hypo (1-)
DFS-SF study 2 – *Dependence /daily life*[34]	English, and other							a. 347	a. Adequate	a. Results in line with 2 hypos (2+); results	c. 252	c. Very good	c. Results in line with 1 hypo (1+)

										not in line with 3 hypos (3-)			
DFS-SF Polish – *Dependence /daily life*[45]	Polish							a. 212 b1. 212 b2. 212	a. Very good b1. Inadequate b2. Doubtful	a. Results in line with 4 hypos (4+); results not in line with 3 hypos (3-) b1. Results in line with 1 hypo (1+); results not in line with 3 hypos (3-) b2. No data provided (?)			
DFS-SF Chinese – *Dependence /daily life*[38]	Hong Kong Chinese							a. 60 b. 60	a. Very good b. Very good	a. Results in line with 5 hypos (5+); results not in			

										line with 2 hypos (2-) b. Results in line with 1 hypo (1+)[8]; results not in line with 1 hypo (1-)			
DFS-SF Greek – *Dependence /daily life*[60]	Greek							a. 110 b. 110[j]	a. Very good b. Very good	a. Results in line with 4 hypos (4+); results not in line with 3 hypos (3-) b. Results in line with 14 hypos (14+); results not in line with 1 hypo (1-)			
DFS-SF	Spanish							a.	a.	a. Results	c. 141	c. Very	c. Results in line

Spanish – *Dependence /daily life*[61]								141	Adequate	in line with 4 hypos (4+); results not in line with 1 hypo (1-)		good	with 1 hypo (1+)
Pooled or summary result (overall rating)								**a. 1050 b. 170**	**a. High b. High**	**a. 21+ and 15- (±) b. 15+ and 2- (+)**	**c. 573**	**c. High**	**c. 2+ and 1- (±)**
PRO-DM-Thai – *Physical function*[35]	Thai							b1. 200 b2. 200	b1. Inadequate b2. Very good	b1. Results not in line with 1 hypo (1-) b2. No data provided (?)			
Pooled or summary result (overall rating)								**b. 200**	**b. Very low**	**b. 1- (-)**			
IWADL/APPADL – *(Physical) activities of daily living*[36]	English							b. 349	b. Very good	b. Results in line with 12 hypos (12+); results not in line with			

										22 hypos (22-)			
IWADL/APPADL – *(Physical) activities of daily living*[46]	English	106	Doubtful	(physical) activities of daily living: SEM=6.3; SDC=17.5; MIC=9.8-13.6[k] (-)							d. 40	d. Very good	d. Results in line with 2 hypos (2+); results not in line with 1 hypo (1-)
Pooled or summary result (overall rating)		**106**	**Low**	**(physical) activities of daily living: SEM=6.3; SDC=17.5; MIC=9.8-13.6[k] (-)**				**b. 349**	**b. High**	**b. 12+ and 22- (±)**	**d. 40**	**d. Low**	**d. 2+ and 1- (±)**
QOLID – *Physical endurance*[37]	English/Hindi[f]							a. 30 b. 210[l]	a. Adequate b. Very good	a. Results in line with 3 hypos (3+); results not in line with 3 hypos (3-) b. Results not in line with 3 hypos (3-)			
Pooled or summary result (overall rating)								**a. 30 b. 210**	**a. Very low b. High**	**a. 3+ and 3 - (±) b. 3-**			

										(-)			
DQLCTQ – *Physical function*[44]	English, German, French							b. 942[m]	b. Very good	b. Results in line with 1 hypo (1+); results not in line with 1 hypo (1-)	c. 328	c. Very good	c. Results not in line with 2 hypos (2-)
Pooled or summary result (overall rating)								**b. 942**	**b. High**	**b. 1+ and 1- (±)**	**c. 328**	**c. High**	**c. 2- (-)**
Diabetes-39 – *Energy and mobility (pilot version - 14 items)*[43]	English							b. 516	b. Adequate	b. No data provided (?)			
Pooled or summary result (overall rating)								**b. 516**		**b. (?)**			
Diabetes-39 – *Energy and mobility (15 items)*[43]	English							a. 427 b. 427	a. Adequate b. Doubtful	a. Results in line with 6 hypos (6+); results not in			

										line with 4 hypos (4-) b. No data provided (?)			
Diabetes-39 Arabic – *Energy and mobility (15 items)*[39]	Arabic							a. 368	a. Adequate	a. Results in line with 2 hypos (2+); results not in line with 3 hypos (3-)			
Diabetes-39 Vietnamese – *Energy and mobility (15 items)*[40]	Vietnamese							b. 286	b. Very good	b. Results in line with 4 hypos (4+); results not in line with 6 hypos (6-)[12]			
Diabetes-39 Portuguese – *Energy and mobility (15 items)*[41]	Portuguese							b. 52	b. Very good	b. No data provided (?)			
Diabetes-39 Spanish – *Energy and*	Spanish							b. 249	b. Very good	b. No data provided			

mobility (15 items)[51]										(?)			
Diabetes-39 Taiwan – *Energy and mobility (15 items)*[59]	Chinese							a. 280 b. 280	a. Very good b. Very good	a. No hypos defined (?) b. Results in line with 7 hypos (7+); results not in line with 5 hypos (5-)			
Diabetes-39 Taiwan – *Energy and mobility (15 items)*[48]	Chinese												
Pooled or summary result (overall rating)								**a. 795** **b. 566**	**a. High** **b. High**	**a. 8+ and 7- (±)** **b. 11+ and 11- (±)**			
Diabetes-39 Thai – *Energy and mobility (10 items)*[49]	Thai							a. 397 b. 397	a. Adequate b. Very good	a. Results in line with 2 hypos (2+); results not in line with			

										3 hypos (3-) b. Results in line with 4 hypos (4+); results not in line with 2 hypo (2-)			
Pooled or summary result (overall rating)								**a. 397 b. 397**	**a. Moderate b. High**	**a. 2+ and 3- (±) b. 4+ and 2- (±)**			
Diabetes-39 German – *Physical impairment (7 items)*[50]	German							a. 144 b. 144	a. Very good b. Very good	a. No hypos defined (?) b. Results in line with 4 hypos (4+)	d. 62-66	d. Doubtful	d. No data provided (?)
Pooled or summary result (overall rating)								**a. 144 b. 144**	**b. High**	**a. (?) b. 4+ (+)**	**d. 62-66**		**d. (?)**
Diabetes-39 SF – *Energy and mobility*	Chinese												

(5 items)[48]													
Pooled or summary result (overall rating)													
C-CWIS – *Physical symptoms and everyday living*[47]	Chinese							a. 131 b. 131	a. Inadequate b. Inadequate	a. No data provided (?) b. Result in line with 1 hypo (1+)			
Pooled or summary result (overall rating)									**b. Very low**	**a. (?) b. 1+ (+)**			

a α refers to Cronbach's alpha; b Four moments of measurement; c Three moments of measurement; d Three moments of measurements; T1: n=119, T2: n=106, T3: n=40; e Per protocol of the COSMIN guideline for systematic reviews: the quality of evidence for internal consistency cannot be higher than the quality of evidence for structural validity[23]; f Language unsure; g Sample size unsure; h Reliability parameter unknown; i One of the known-groups tested in the hypotheses was small (n=10); j Some of the known-groups tested in the hypotheses were small (n<10); k MIC based on ability to perform daily physical activities and MIC based on weight loss, respectively; 10 n=60 for HbA1C levels; l n=274 for HbA1C levels; m One of the known-groups tested in the hypotheses was small (n=14)

Appendix 11. Extensive summary of findings

STRUCTURAL VALIDITY			
PROM – *subscale*	**Summary or pooled result**	**Overall rating**	**Quality of evidence**
DFS – *Daily activities*[42]	Two subscales: Daily life (3 items) and Dependence (4 items)	Indeterminate	No level of evidence: rating is indeterminate
DFS-SF – *Dependence/daily life*[34,38,45,60,61]	Unidimensional scale	Sufficient	High: one very good study (results of inadequate study are ignored)
PRO-DM-Thai – *Physical function*[35]	Unidimensional scale	Sufficient	High: one very good study
IWADL/ APPADL – *(Physical) activities of daily living*[36,46]	Unidimensional scale	Sufficient	Low: one doubtful study
QOLID – *Physical endurance*[37]	Unidimensional scale	Sufficient	Very low: one inadequate study
DQLCTQ – *Physical function*[44]			
Diabetes-39 – *Energy and mobility (pilot version - 14 items)*[43]	Unidimensional scale	Sufficient	Moderate: one adequate study
Diabetes-39 – *Energy and mobility (15 items)*[39-41,43,48,51,59]	Unidimensional scale	Sufficient	High: two adequate studies
Diabetes-39 Thai – *Energy and mobility (10 items)*[49]	Structural validity unconfirmed	Insufficient	Moderate: one adequate study
Diabetes-39 German – *Physical impairment (7 items)*[50]	Structural validity unconfirmed	Indeterminate	No level of evidence: rating is indeterminate
Diabetes-39 SF – *Energy and mobility (5 items)*[48]	Unidimensional scale	Sufficient	High: one very good study
C-CWIS – *Physical symptoms and everyday living*[47]	Unidimensional scale	Sufficient	Moderate: one adequate study

INTERNAL CONSISTENCY			
PROM – *subscale*	**Summary or pooled result**	**Overall rating**	**Quality of evidence**
DFS – *Daily activities*[42]	Cronbach's alpha: 0.85; n=48	Indeterminate	No level of evidence: rating is indeterminate because structural validity is

			indeterminate
DFS-SF – *Dependence/daily life*[34,38,45,60,61]	Cronbach's alpha: 0.85-0.91; consistent results; n=1050	Sufficient	High: six very good studies
PRO-DM-Thai – *Physical function*[35]	Cronbach's alpha: 0.82; n=500	Sufficient	High: one very good study
IWADL/ APPADL – *(Physical) activities of daily living*[36,46]	Cronbach's alpha: ≥0.89; consistent results; n=455;	Sufficient	Low: two very good studies, but structural validity is low
QOLID – *Physical endurance*[37]	Cronbach's alpha: 0.85; n=150	Indeterminate	No level of evidence: rating is indeterminate because structural validity is very low
DQLCTQ – *Physical function*[44]	Cronbach's alpha: 0.85; n=909	Indeterminate	No level of evidence: rating is indeterminate because structural validity is not assessed
Diabetes-39 – *Energy and mobility (pilot version - 14 items)*[43]	Cronbach's alpha: 0.81-0.92; n=516	Sufficient	Moderate: one very good study, but structural validity is moderate
Diabetes-39 – *Energy and mobility (15 items)*[39-41,43,48,51,59]	Cronbach's alpha: 0.79-0.93; consistent results; n=1622	Sufficient	High: six very good studies
Diabetes-39 Thai – *Energy and mobility (10 items)*[49]	Cronbach's alpha: 0.94; n=397	Indeterminate	No level of evidence: rating is indeterminate because structural validity is insufficient
Diabetes-39 German – *Physical impairment (7 items)*[50]	Cronbach's alpha: 0.84; n=144	Indeterminate	No level of evidence: rating is indeterminate because structural validity is indeterminate
Diabetes-39 SF – *Energy and mobility (5 items)*[48]			
C-CWIS – *Physical symptoms and everyday living*[47]	Cronbach's alpha: 0.73-0.92; n=151	Sufficient	Moderate: one very good study, but structural validity is moderate

CROSS-CULTURAL VALIDITY\MEASUREMENT INVARIANCE			
PROM – *subscale*	**Summary or pooled result**	**Overall rating**	**Quality of evidence**

DFS – *Daily activities*[42]			
DFS-SF – *Dependence/daily life*[34,38,45,60,61]	DIF for gender in one item, while five items were tested for six variables; n=212	Sufficient	Very low: one inadequate study
PRO-DM-Thai – *Physical function*[35]			
IWADL/ APPADL – *(Physical) activities of daily living*[36,46]			
QOLID – *Physical endurance*[37]			
DQLCTQ – *Physical function*[44]			
Diabetes-39 – *Energy and mobility (pilot version - 14 items)*[43]			
Diabetes-39 – *Energy and mobility (15 items)*[39-41,43,48,51,59]			
Diabetes-39 Thai – *Energy and mobility (10 items)*[49]			
Diabetes-39 German – *Physical impairment (7 items)*[50]			
Diabetes-39 SF – *Energy and mobility (5 items)*[48]			
C-CWIS – *Physical symptoms and everyday living*[47]			

RELIABILITY			
PROM – *subscale*	**Summary or pooled result**	**Overall rating**	**Quality of evidence**
DFS – *Daily activities*[42]	Daily life: r=0.68; n=? Dependence: r=0.84; n=?	Daily life: insufficient Dependence: sufficient	Very low: one inadequate study
DFS-SF – *Dependence/daily life*[34,38,45,60,61]	ICC: 0.74-0.77; consistent results; n>141	Sufficient	Low: one doubtful study
PRO-DM-Thai – *Physical function*[35]			
IWADL/ APPADL – *(Physical) activities of daily living*[36,46]	ICC: 0.91; n=106	Sufficient	Moderate: one adequate study

QOLID – *Physical endurance*[37]	ICC: 0.83; n=58	Sufficient	Low: one adequate study, sample size 58
DQLCTQ – *Physical function*[44]			
Diabetes-39 – *Energy and mobility (pilot version - 14 items)*[43]			
Diabetes-39 – *Energy and mobility (15 items)*[39-41,43,48,51,59]	ICC: 0.91, spearman r: 0.84; consistent results; n=535	Sufficient	Moderate: two doubtful studies
Diabetes-39 Thai – *Energy and mobility (10 items)*[49]			
Diabetes-39 German – *Physical impairment (7 items)*[50]	0.88; n=72	Indeterminate	No level of evidence: rating is indeterminate
Diabetes-39 SF – *Energy and mobility (5 items)*[48]			
C-CWIS – *Physical symptoms and everyday living*[47]			

MEASUREMENT ERROR			
PROM – *subscale*	**Summary or pooled result**	**Overall rating**	**Quality of evidence**
DFS – *Daily activities*[42]			
DFS-SF – *Dependence/daily life*[34,38,45,60,61]			
PRO-DM-Thai – *Physical function*[35]			
IWADL/ APPADL – *(Physical) activities of daily living*[36,46]	SEM: 6.3, SDC: 17.5, MIC: 9.8-13.6; n=106	Insufficient	Low: one doubtful study
QOLID – *Physical endurance*[37]			
DQLCTQ – *Physical function*[44]			
Diabetes-39 – *Energy and mobility (pilot version - 14 items)*[43]			
Diabetes-39 – *Energy and mobility (15 items)*[39-41,43,48,51,59]			
Diabetes-39 Thai – *Energy and mobility (10 items)*[49]			

Diabetes-39 German – *Physical impairment (7 items)*[50]			
Diabetes-39 SF – *Energy and mobility (5 items)*[48]			
C-CWIS – *Physical symptoms and everyday living*[47]			

HYPOTHESES TESTING FOR CONSTRUCT VALIDITY **a=comparison with other instruments** **b=comparison between subgroups**			
PROM – *subscale*	**Summary or pooled result**	**Overall rating**	**Quality of evidence**
DFS – *Daily activities*[42]	a. 4 out of 5 hypotheses confirmed; n=173 b. No data provided	a. Sufficient b. Indeterminate	a. Moderate: one adequate study b. No level of evidence: rating is indeterminate
DFS-SF – *Dependence/daily life*[34,38,45,60,61]	a. 21 out of 36 hypotheses confirmed; n=1050 b. 15 out of 17 hypotheses confirmed; n=170	a. Inconsistent b. Sufficient	a. High: three very good studies b. High: two very good studies (results of inadequate study and study with no data provided are ignored)
PRO-DM-Thai – *Physical function*[35]	b. 0 out of 1 hypothesis confirmed; n=200	b. Insufficient	b. Very low: one inadequate study (results of study with no data provided are ignored)
IWADL/ APPADL – *(Physical) activities of daily living*[36,46]	b. 12 out of 34 hypotheses confirmed; n=349	b. Inconsistent	b. High: one very good study
QOLID – *Physical endurance*[37]	a. 3 out of 6 hypotheses confirmed; n=30 b. 0 out of 3 hypotheses confirmed; n=210	a. Inconsistent b. Insufficient	a. Very low: one adequate study, sample size 30 b. High: one very good study
DQLCTQ – *Physical function*[44]	b. 1 out of 2 hypotheses confirmed; n=942	b. Inconsistent	b. High: one very good study
Diabetes-39 – *Energy and mobility (pilot version - 14 items)*[43]	b. No data provided	b. Indeterminate	b. No level of evidence: rating is indeterminate

Diabetes-39 – *Energy and mobility (15 items)*[39-41,43,48,51,59]	a. 8 out of 15 hypotheses confirmed; n=795 b. 11 out of 22 hypotheses confirmed; n=566	a. Inconsistent b. Inconsistent	a. High: two adequate studies (results of study with no data provided are ignored) b. High: two very good studies (results of study with no data provided are ignored)
Diabetes-39 Thai – *Energy and mobility (10 items)*[49]	a. 2 out of 5 hypotheses confirmed; n=397 b. 4 out of 6 hypotheses confirmed; n=397	a. Inconsistent b. Inconsistent	a. Moderate: one adequate study b. High: one very good study
Diabetes-39 German – *Physical impairment (7 items)*[50]	a. No data provided b. 4 out of 4 hypotheses confirmed; n=144	a. Indeterminate b. Sufficient	a. No level of evidence: rating is indeterminate b. High: one very good study
Diabetes-39 SF – *Energy and mobility (5 items)*[48]			
C-CWIS – *Physical symptoms and everyday living*[47]	a. No data provided b. 1 out of 1 hypotheses confirmed; n=131	a. Indeterminate b. Sufficient	a. No level of evidence: rating is indeterminate b. Very low: one inadequate study

RESPONSIVENESS **a=comparison to gold standard** **b=comparison with other instruments** **c=comparison between subgroups** **d=before and after intervention**			
PROM – *subscale*	**Summary or pooled result**	**Overall rating**	**Quality of evidence**
DFS – *Daily activities*[42]	c. 0 out of 2 hypotheses confirmed; n=264	c. Insufficient	c. Low: one doubtful study
DFS-SF – *Dependence/daily life*[34,38,45,60,61]	c. 2 out of 3 hypotheses confirmed; n=573	c. Inconsistent	c. High: three very good studies
PRO-DM-Thai – *Physical function*[35]			
IWADL/ APPADL – *(Physical) activities of daily living*[36,46]	d. 2 out of 3 hypotheses confirmed; n=40	d. Inconsistent	d. Low: one very good study, sample size 40

QOLID – *Physical endurance*[37]			
DQLCTQ – *Physical function*[44]	c. 0 out of 2 hypotheses confirmed; n=328	c. Insufficient	c. High: one very good study
Diabetes-39 – *Energy and mobility (pilot version - 14 items)*[43]			
Diabetes-39 – *Energy and mobility (15 items)*[39-41,43,48,51,59]			
Diabetes-39 Thai – *Energy and mobility (10 items)*[49]			
Diabetes-39 German – *Physical impairment (7 items)*[50]	d. No data provided	d. Indeterminate	d. No level of evidence: rating is indeterminate
Diabetes-39 SF – *Energy and mobility (5 items)*[48]			
C-CWIS – *Physical symptoms and everyday living*[47]			

Appendix 12. Quality of PROM development and content validity

PROM	PROM design							Cognitive interview (CI) study[a]				Total PROM development	Content validity				
	General design requirements					Concept elicitation	Total PROM design	General design requirements	Comprehensibility	Comprehensiveness	Total CI study		Asking patients			Asking experts	
	Clear construct	Clear origin of construct	Clear target population for which the PROM was developed	Clear context of use	PROM developed in sample representing the target population			CI study performed in sample representting the target population					Relevance	Comprehensivness	Comprhensibility	Relevance	Comprehensiveness
DFS[42]	I	D	V	V	A	D	I	A	D		D	I					
DFS-SF[34,38,61]	V	D	V	V	A	D	D	A	D			D	I[b]		D[b]		
PRO-DM-Thai[35]	V	D	V	D	D	D	D	A	D		D	D					
IWADL/ APPADL[36]	V	D	V	V	V	D	D	V	?	?	D	D					
QOLID[37]	V	D	V	V	A	D	D	D	?	?	D	D					
DQLCTQ[44]	I	V	V	V	D	D	I					I					
Diabetes-39[39-41,43]	I	V	V	D	D	D	I					I			D[c]		
Diabetes-39 SF[48]	I	V	V	D	D	D	I					I					
C-CWIS[47]	I	D	V	V	I		I	A	D		D	I					

a Empty cells indicate that a CI study or content validity study (or part of it) was not performed, ? indicates that something was done but unclear what was done; b Content validity of the Chinese and Spanish version were assessed; c Content validity of the Arabic, Vietnamese and Portuguese version were assessed