REVIEW

Measurement properties of the EQ-5D in populations with a mean age of ≥ 75 years: a systematic review

Sophie Gottschalk[1] · Hans-Helmut König[1] · Mona Nejad[1] · Judith Dams[1]

Abstract

Purpose Healthcare interventions for middle-old and oldest-old individuals are often (economically) evaluated using the EQ-5D to measure health-related quality of life (HrQoL). This requires sufficient measurement properties of the EQ-5D. Therefore, the current study aimed to systematically review studies assessing the measurement properties of the EQ-5D in this population.

Methods The databases PubMed, Cochrane library, Web of Science, Embase, and EconLit were searched for studies providing empirical evidence of reliability, validity, and/or responsiveness of the EQ-5D-3L and EQ-5D-5L in samples with a mean age ≥ 75 years. Studies were selected by two independent reviewers, and the methodological quality was assessed using the COSMIN Risk of Bias checklist. Results were rated against updated criteria for good measurement properties (sufficient, insufficient, inconsistent, indeterminate). The evidence was summarized, and the quality of evidence was graded using a modified GRADE approach.

Results For both EQ-5D versions, high-quality evidence for sufficient convergent validity was found. Known-groups validity was sufficient for the EQ-5D-5L (high-quality evidence), whereas the results were inconsistent for the EQ-5D-3L. Results regarding the reliability were inconsistent (EQ-5D-3L) or entirely lacking (EQ-5D-5L). Responsiveness based on correlations of change scores with instruments measuring related/similar constructs was insufficient for the EQ-5D-3L (high-quality evidence). For the EQ-5D-5L, the available evidence on responsiveness to change in (Hr)QoL instruments was limited.

Conclusion Since the responsiveness of the EQ-5D in a population of middle-old and oldest-old individuals was questionable, either using additional instruments or considering the use of an alternative, more comprehensive instrument of (Hr)QoL might be advisable, especially for economic evaluations.

Keywords EQ-5D · Older population · Oldest-old population · Psychometric properties · Systematic review

Abbreviations

ADL	Activities of daily living
ASCOT	Adult Social Care Outcomes Toolkit
AQoL	Assessment of Quality of Life
BBS	Berg Balance Scale
BADL	Bristol Activities of Daily Living Scale
CCCQ	Client-centered Care Questionnaire
CDR	Clinical Dementia Rating
CMAI	Cohen-Mansfield Agitation Inventory
COSMIN	COnsensus-based Standards for the selection of health Measurement INstruments
DEMQOL	Dementia Quality of Life instrument
ESM	Electronic supplementary material
EQ-HWB	EQ Health and Wellbeing instrument
EQ-VAS	EQ-Visual Analogue Scale
FAST	Functional Assessment Staging Tool
HrQoL	Health-related quality of life
HUI3	Health Utilities Index
IADL	Instrumental activities of daily living
ICC	Intraclass correlation coefficient

✉ Sophie Gottschalk
s.gottschalk@uke.de

Hans-Helmut König
h.koenig@uke.de

Mona Nejad
m.nejad@uke.de

Judith Dams
j.dams@uke.de

[1] Department of Health Economics and Health Services Research, University Medical Center Hamburg-Eppendorf, Martinistraße 52, 20246 Hamburg, Germany

ICECAP-O	ICEpop CAPability measure for Older people
MBI	Modified Barthel Index
MCID	Minimally clinically important difference
MeSH	Medical subject headings
NHP	Nottingham Health Profile
NOSGER	Nurses' Observation Scale for Geriatric Patients
OHS	Oxford Hip Score
OPQOL-Brief	Older People's Quality of Life questionnaire, short version
PPA	Physiological Profile Assessment
PRISMA	Preferred Reporting Items for Systematic reviews and Meta-Analysis
PROMIS	Patient-Reported Outcomes Measurement Information System
QALY	Quality-adjusted life years
QoL	Quality of life
QoL-AD	Quality of Life in Alzheimer's Disease scale
QOL-AD-NH	Quality of Life in Alzheimer's Disease in Nursing Homes
QWB	Quality of Well-Being scale
SF-36	36-item Short-Form health survey
SF-12	12-item Short-Form health survey
SF-6D	Short Form 6 Dimensions
SPPB	Short Physical Performance Battery
SPVU-5D	5-Dimensional Sheffield Preference-based Venous Ulcer questionnaire
UK	United Kingdom
US	United States
WHOQOL-OLD	World Health Organization Quality of Life - Older Adults
30 s STS	30-second Sit-To-Stand test

Introduction

Maintaining health of an increasing number of middle-old and oldest-old people is a major challenge for aging societies [1]. Population norms of health-related quality of life (HrQoL) suggest that HrQoL decreases with age and drops considerably beyond the age of 75 [2, 3]. Numerous interventions targeting this population are, therefore, being developed. In the face of scarce resources, new interventions should be economically evaluated before being implemented in the healthcare system, as such information can assist in the efficient allocation of resources.

To make effects comparable across interventions, economic evaluations often measure effectiveness in terms of quality-adjusted life years (QALY), where the 'Q' is measured using generic HrQoL instruments. The most frequently

used instrument, in general but also for evaluation of interventions targeting the older population, is the EQ-5D [4–6], which is the officially required standard measurement in some countries (e.g., UK [7]). It consists of five questions covering the dimensions mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Depending on the version of the EQ-5D, each dimension has three (EQ-5D-3L) or five (EQ-5D-5L) severity levels ("no problems" to "extreme problems"). The combined answers can be transformed to an index with 0 representing death and 1 representing the best possible HrQoL. It is important that the EQ-5D is psychometrically sound in the population it is used, meaning that it measures what it intended to measure (validity) in an accurate and reproducible way (reliability) and is able to detect important changes over time (responsiveness). In the absence of sufficient measurement properties, the results of economic evaluations fail in measuring the true effect of interventions and, thus, are not suitable as basis for decision making regarding their implementation.

Previous reviews examined the psychometric performance of the EQ-5D in different population groups. It was found appropriate for depression and personality disorders [8, 9], urinary incontinence [10], some skin diseases [11], and in people aged 60 or older [12]. However, its psychometric performance was lacking in populations with anxiety, schizophrenia, bipolar disorders, or multiple sclerosis [8, 9, 13]. Moreover, it was found insufficiently sensitive to change in a range of disorders [14]. Regarding its use in dementia, the validity was found problematic as there are significant disagreements between patient and proxy ratings and aspects being important for people with dementia are not adequately reflected [15, 16]. Similarly, other authors conclude that the EQ-5D may not be appropriate in other conditions prevalent in the older population, such as hearing impairments, visual disorders, and some cancers [17, 18]. A common problem seems to be that the EQ-5D has limited ability to differentiate between healthier individuals [19]. Although this ceiling effect could be reduced for the EQ-5D-5L, it still exists [20]. Moreover, the EQ-5D has been criticized for its narrow focus of health, which may fall short on or excludes important aspects of health (e.g., social aspects) [21]. As people's needs and desires change with age, it can be assumed that, especially in old age or at the end of life, such aspects become more important [22–24].

These findings raise questions regarding the measurement properties of the EQ-5D in middle-old and oldest-old people. To our knowledge, there has been no systematic summary of the measurement properties of the EQ-5D in this population. In a review that is more than a decade old, Haywood et al. [12] evaluated the measurement and practical properties of generic health instruments in older people and found evidence for the validity of the EQ-5D. In terms of responsiveness, the EQ-5D appeared to perform well in

people with substantial changes in health; however, responsiveness in terms of correlation of change scores between the EQ-5D and other (clinical) measures was rarely addressed until then. In addition to being outdated and hence including only studies using the EQ-5D-3L, this review did not specifically focus on middle-old and oldest-old people. More recent reviews concluded that the EQ-5D has good feasibility properties in an older population [25], but due to its sole focus on health status, may not be appropriate for measuring outcomes in economic evaluation within aged care, especially in interventions that have effects beyond health status [6, 26, 27]. However, the authors focused exclusively on dependent older people and/or did not systematically summarize the measurement properties of the EQ-5D. Therefore, the aim of the current study was to extend the existing literature by synthesizing and critically appraising studies assessing the measurement properties—reliability, validity, or responsiveness—of the EQ-5D in a population of middle-old and oldest-old people (mean age ≥ 75 years).

Materials and methods

This review was conducted in adherence with the Consensus-Based Standards for the Selection of Health Measurement Instrument (COSMIN) Methodology for Systematic Reviews of Measurement Properties of PROMs [28]. It has been registered with PROSPERO (Registration Number: CRD42020196070), and a study protocol has been published [29]. The manuscript was prepared based on the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) checklist (electronic supplementary material [ESM] 1) [30].

Eligibility criteria

Cross-sectional or observational studies providing empirical evidence of reliability, validity, and/or responsiveness of the EQ-5D in a sample with a mean age of ≥ 75 years were included. Studies had to be published in peer-reviewed journals in German or English languages. Systematic reviews, studies applying a qualitative design, or not being original research articles (e.g., conference abstracts or comments) were excluded. Furthermore, studies relying on proxy assessments only or those with the single objective of investigating agreement between different modes of administration of the EQ-5D were excluded. The question of inter-rater agreement between the patient and a proxy often concerns people with dementia and has been addressed in previous reviews [15, 16]. No restrictions relating to interventions, health conditions, publication date, or the version of the EQ-5D (3-level or 5-level) were made.

Data sources and search strategy

PubMed, Web of Science, Cochrane Library, Embase, and EconLit were searched electronically on March 10, 2021 using predefined search terms, including *quality of life*, *health-related quality of life*, *EQ-5D*, *EuroQoL*, *aged*, *elder**, *old**, *geriatric**, and *ag(e)ing* and an adapted search filter for finding studies on measurement properties [31]. Search terms covering non-relevant measurement properties were removed from the search filter (e.g., inter-rater reliability or cross-cultural validity). Where possible, search terms were used as keywords in the title/abstract or Medical Subject Headings (MeSH). An example for the search strategy in PubMed is displayed in Table S1 (ESM 1). Additionally, reference lists of included studies were hand searched.

Selection of studies and data extraction

Search results from all databases were combined in a shared data repository and managed with Endnote X8. After removing duplicates, two independent reviewers (SG and MN) screened the titles and abstracts and assessed the full texts of the selected abstracts for eligibility. In case of disagreement or uncertainty, a third person (JD) was consulted. Using a standardized data extraction sheet, relevant data from the eligible studies were extracted by one reviewer (SG) and cross-checked by the second reviewer (MN). Data extracted from the individual studies included setting/country, population characteristics, type and method of validity, reliability and responsiveness assessment, and results for each measurement property.

Assessment of study quality

Methodological quality of included studies was assessed by two reviewers (MN and SG) using the COSMIN Risk of Bias checklist, which was developed specifically for the use in systematic reviews of patient-reported outcome measures [32]. It consists of 10 boxes, each referring to a particular measurement property and containing a different number of sub-questions. Each item is rated on a four-point scale ("very good" to "inadequate"). Any disagreements were resolved through discussion with a third person (JD). Risk of bias rating for each study and measurement property are provided in ESM 2.

Evaluation of measurement properties

Updated criteria for good measurement properties were applied to rate the individual studies' results as "sufficient" (+), "insufficient" (−), or "indeterminate" (?) [33]. Reliability was considered "sufficient" if the intraclass correlation coefficient (ICC) was ≥ 0.70. Construct validity and

Table 1 Generic/general hypotheses for construct validity and responsiveness (adapted from Prinsen et al., [28])

H1	Correlations with (changes in) instruments measuring similar constructs should be high (≥ 0.5)
H2	Correlations with (changes in) instruments measuring related, but dissimilar constructs should be at least moderate (≥ 0.3)
H3	Correlations with (changes in) instruments measuring weakly related constructs should be at least weak (≥ 0.1)
H4	Correlations with (changes in) instruments measuring unrelated constructs should be negligible (< 0.1)
H5	Meaningful changes between relevant (sub) groups. MCID of the EQ-5D: 0.074 (EQ-5D-3L)[35] or 0.063 (EQ-5D-5L) [36]
H6	For responsiveness (criterion approach), AUC should be ≥ 0.7
H7	HrQoL may decreases with age, but not necessarily, given the circumstances that this review focusses only on middle-old to oldest-old people
H8	Higher education level/social class might be associated with higher HrQoL, but not necessarily, since the differences may no longer be present in this age group (in later life, lifestyle factors such as physical activity become more important [103])
H9	Lower cognitive status is hypothesized to be associated with lower HrQoL in institutionalized people and/or people with severe dementia, whereas this association may not be visible in people with mild to moderate dementia or non-institutionalized people [102]

AUC area under the curve, *MCID* minimal clinically important differences

responsiveness were rated "sufficient" if the result was in accordance with predefined hypotheses. The hypotheses were formulated by the review team in advance and where partly (but not necessarily) adopted from the authors of the individual studies. Generic hypotheses applied in this study are presented in Table 1. A detailed overview of specific hypotheses for each individual study is provided in Table S2, ESM 1. The hypotheses regarding the discriminative ability of the EQ-5D between relevant subgroups (e.g., known-groups validity or responsiveness) were accepted if the difference between subgroups was clinically relevant, which was considered more important than whether the difference is statistically significant [34]. For the EQ-5D-3L index, a minimally clinically important difference (MCID) of 0.074 was applied, which was identified as the mean MCID across different patient groups [35]. The studies reporting on known-groups validity or responsiveness of the EQ-5D-5L index were either conducted in the UK or used UK value sets. Therefore, an MCID of 0.063 was applied, which was identified as MCID for England [36].

Summary and grading of the quality of evidence

Criteria for good measurement properties were applied to the summarized results from the individual studies on each measurement property by rating each property as "sufficient" (+), "insufficient" (−), "inconsistent" (±), or "indeterminate" (?) [33, 37]. For construct validity and responsiveness, the measurement property was rated "sufficient" when $\geq 75\%$ of the individual studies' results were in accordance with predefined hypotheses. The results were qualitatively summarized by providing, e.g., a range of correlation coefficients for convergent validity and the percentage of hypotheses accepted. The evidence synthesis was performed separately for the EQ-5D-3L and EQ-5D-5L. If the results were inconsistent, reasons for inconsistency were explored

(e.g., different results for different subgroups). If no reason for inconsistency could be identified, the result was rated "inconsistent" and the quality of evidence was not further explored. Due to heterogeneity of the populations included in the individual studies, quantitative pooling of results was not performed.

The quality of evidence was graded as "high," "moderate," "low," or "very low" using a modified GRADE approach [38]. Starting with the assumption of "high quality," it was downgraded if there was a risk of bias (up to − 3 levels), (unexplained) inconsistency (up to − 2 levels), imprecision (e.g., small sample size; up to − 2 levels), or indirect results. Indirectness was not applied in this study since studies examining the measurement properties in other populations than the population of interest were excluded. Specific criteria for downgrading are described in the COSMIN manual [34].

Results

Search results

The search strategy resulted in 4346 records (duplicates removed). After screening of title and abstract, 4107 records were excluded, leaving 239 records of which full texts were assessed for eligibility. Finally, 38 records were included for the qualitative synthesis (Fig. 1). No further relevant studies were identified through reference screening. The majority of studies ($n = 30$) evaluated the measurement properties of the EQ-5D-3L [39–68], whereas 9 studies evaluated the EQ-5D-5L [41, 69–76]. One study evaluated both EQ-5D versions [41].
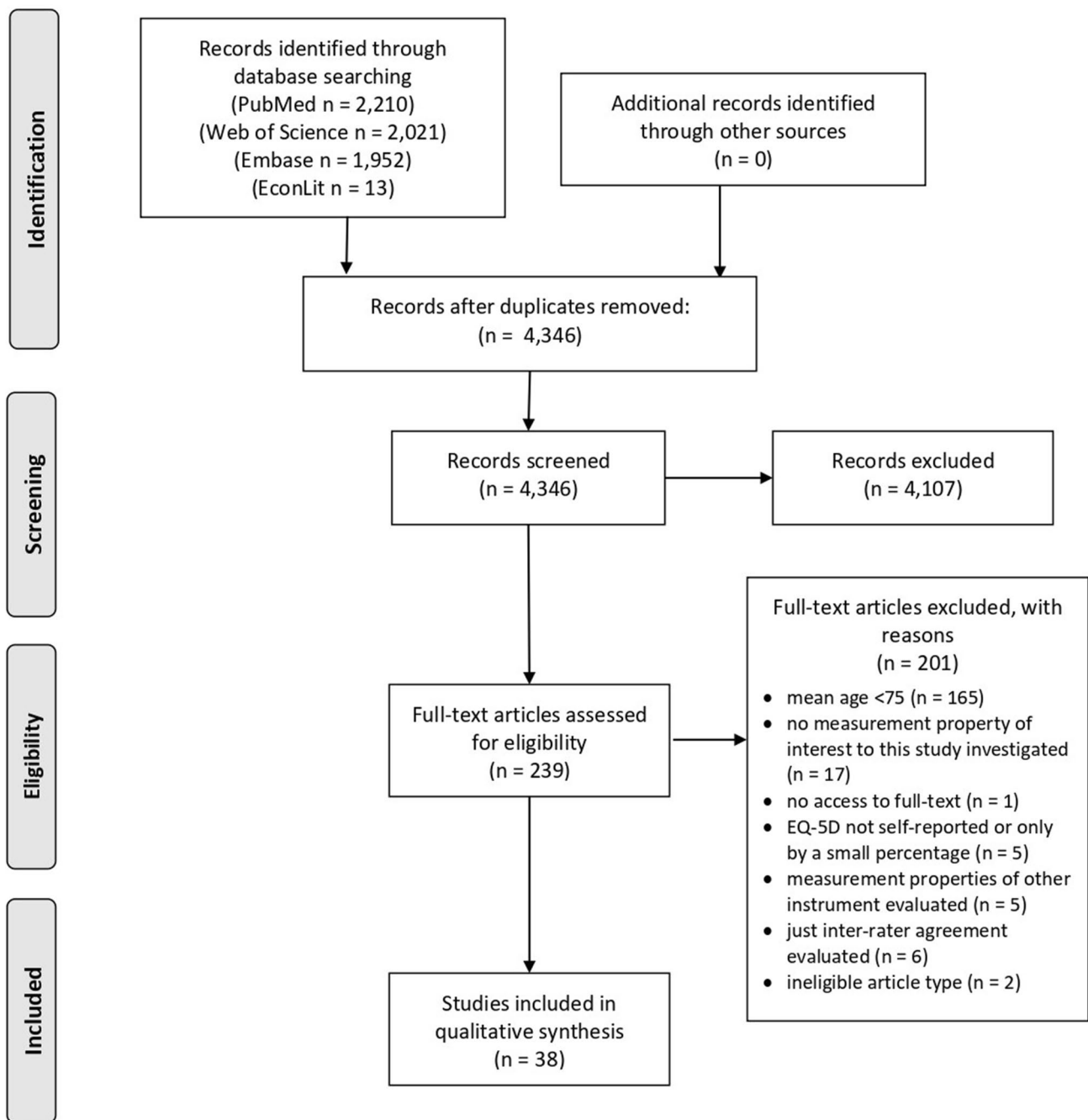
Fig. 1 Selection process of included studies

General characteristics of the articles

Characteristics of the included studies are described in Table 2. Studies covered a variety of (disease) populations, such as people with dementia or cognitive impairment (*n* = 13) [39, 50, 52, 54, 57, 58, 60, 62–64, 69, 72–74], people with different kinds of fractures (*n* = 7) [43, 46, 59, 61, 65, 66, 76], people who were frail or had a history of falling (*n* = 4) [44, 45, 67, 70], or people with venous leg

ulcers (*n* = 2) [68, 71]. The studies were conducted in the UK (*n* = 12) [40, 42, 43, 47, 49, 60, 61, 68, 69, 73–75], Sweden (*n* = 3) [59, 65, 66], Spain (*n* = 2) [62, 63], Norway (*n* = 2) [46, 70], Finland (*n* = 1) [48], France (*n* = 1) [39], Germany (*n* = 2) [54, 57], Korea (*n* = 1) [53], the Netherlands (*n* = 2) [55, 67], Australia (*n* = 4) [51, 71, 72, 76], Canada (*n* = 3) [44, 45, 58], the USA (*n* = 2) [52, 56], Mexico (*n* = 1) [64], Sweden/Denmark/Finland/Norway (*n* = 1) [50], or Belgium/Ireland/Netherlands/Switzerland (*n* = 1) [41]. Participants

Table 2 Characteristics of the included studies

Ref	Population			Disease/population characteristics			Instrumental administration	
	N*	Age Mean (SD, range)	% female	Disease/other characteristics of the study population/recruited from	Disease duration	Disease severity	Interview administration mode	Country
EQ-5D-3L								
Ankri et al. [39]	142	82.9 (8.3) (60–99)	79.6%	PwD; hospitalized, institutionalized, or outpatients; recruited from geriatric hospital centers	N.R	47.1% moderate, 27.9% severe	Assisted interviews	France
Barton et al. [40]	392	N.R.[1]	N.R	Registered in one general practice	N.A	N.A	N.R	UK
Brazier et al. [42]	377	80.1 (4.5)	100%	Older women, 86.5% long-standing illness or disability; recruited from four general practices into a RCT of clodronate	N.A	N.A	N.R	UK
Coast et al. [43]	214	79[2] (74–84)[3]	70%	Elderly acute care patients being suitable for rehabilitation in their own home (mainly fractured neck of the femur, elective hip and knee replacements, and other fractures, and stroke)	N.A	N.A	N.R	UK
Davis et al. [44]	215	79.3 (6.2)	71.6%	Older adults at risk of mobility impairment and a fall history; visiting the Vancouver falls prevention clinic	N.A	N.A	N.R	Canada
Davis et al. [45]	356	82 (6.5)	63%	Older adults at risk of mobility impairment and a fall history; visiting the Vancouver falls prevention clinic	N.A	N.A	N.R	Canada

Table 2 (continued)

Ref	Population			Disease/population characteristics			Instrumental administration	
	N*	Age Mean (SD, range)	% female	Disease/other characteristics of the study population/recruited from	Disease duration	Disease severity	Interview administration mode	Country
Frihagen et al. [46]	222 (complete cases at follow-up: n=79)	82.8 (7.48)	74%	Patients with displaced femoral neck fracture	N.R	n=23 complication group, n=56 non-complication group	Outpatient clinic and home visits	Norway
Hazell et al. [47]	721	N.R.[1]	N.R	Registered in two general practices	N.A	N.A	Postal questionnaire	UK
Heiskanen et al. [48]	36	N.R.[1]	N.R	Patients admitted for CABG operation	N.A	>60% Canadian cardiovascular society class 3 or 4 (indicating lower functional status)	As part of the pre-operative hospital admission process (baseline) and via postal survey (follow-up)	Finland
Holland et al. [49]	145	84.7	57%	People taking ≥2 medications; admitted to hospital as an emergency; to be discharged from hospital and returning home/to a warden control accommodation	N.A	Median of 6 daily drugs	Assisted (baseline) interviews	UK
Jönsson et al. [50]	272	75.9	62.3	Pw/AD, recruited among patients attending regular visits at memory clinics; community dwelling or in residential care	Diagnosed on average 1.6 years prior to inclusion	MMSE>25 to MMSE<10	N.R	Sweden, Denmark, Finland, Norway
Kaambwa et al. [51]	87	80 (65–93)	66%	Receiving community aged care services, but cognitively intact	N.A	N.A	Group setting in central venues (research team just there for clarification of questions)	Australia

Table 2 (continued)

Ref	Population N*	Age Mean (SD, range)	Disease/population characteristics				Instrumental administration	
			% female	Disease/other characteristics of the study population/recruited from	Disease duration	Disease severity	Interview administration mode	Country
Karlawish et al. [52]	93	76.8 (2.7) (55–91)	45%	PwAD, not living in care homes, receiving CG assistance; recruited from geriatric medicine practice or memory clinic	N.R	Mild to moderate AD (71% very mild)	Assisted interviews at the participant's/CG's home or other convenient location	USA
Kim et al. [53]	2826	N.R.[1]	0%	General male adult population sample, participating in Korean community health survey	N.A	57% moderate to severe lower urinary tract symptoms	Face-to-face interviews	
Kunz [54]	390	80.2 (6.7) (65–100)	68%	PwD, living at home and supported by a family caregiver; recruited via general practitioners	N.R	Mild to moderate	Assessed at GP practices by trained GP and medical secretary	Germany
Lutomski et al. [55]	25,637	78 (6)	58.3%	Community-dwelling older persons aged 65+; recruited from primary care centers, hospitals, or the general population	N.A	73% with hearing issues, joint damage, urinary incontinence, and/or dizziness with falls	N.R	Netherlands
Malkin et al. [56]	77	77[2] (27–98)	71%	Low-vision patients; presenting for low-vision rehabilitation at private outpatient centers	N.R	55% macular disorders	Telephone interview before 1st visit at low vision rehabilitation site	USA
Michalowsky et al. [57]	560	79.03 (8.5)	59.1%	PwD, living in the community; supported by informal CG	N.R	N.R	Face-to-face interview	Germany
Naglie et al. [58]	57	78.6 (53.8–93.7)	61.7%	PwAD, supported by a family CG; recruited from dementia clinics & geriatric practices	N.R	Mild to moderate	Assisted interview at participant's home/referring clinic	Canada

Table 2 (continued)

Ref	Population			Disease/population characteristics			Instrumental administration	
	N*	Age Mean (SD, range)	% female	Disease/other characteristics of the study population/recruited from	Disease duration	Disease severity	Interview administration mode	Country
Olerud et al. [59]	145	74.7 (9.6) (55–93)	84%	Patients with acute proximal humeral fracture; living non-institutionalized; no severe cognitive dysfunction	N.A	2-part to 4-part fractures	N.R	Sweden
Orgeta et al. [60]	478	75.5 (7.3)	49.6%	PwD living in the community; supported by carers assisting with ADL	N.R	Mild (74.6%) to moderate (25.4%) dementia	Assisted interview	UK
Parsons et al. [61]	225 (sample 1) 249 (sample 2)	83.1 (7.94) 83.6 (7.77)	71% 75%	Patients with hip fracture	N.A	32% and 41% PwD	Face to face at BL; telephone interview at 4 week, 4 month & 12 month FU	UK
Pérez-Ros and Martínez-Arnau [62]	251	84.6 (9.22) (70–104)	76.9%	Nursing home residents with cognitive impairment	N.A	Mean MMSE score: 15.6 (5.23)	Face-to-face interview	Spain
Pérez-Ros et al. [63]	188	79.19 (5.18) (70–95)	64.9%	Community-dwelling older people with cognitive impairment	N.A	MMSE scores 10–24	Face-to-face interview	Spain
Sanchez-Arenas et al. [64]	109	78.5 (7.09)	64.2%	PwD; community dwelling	N.A	N.R	In-home face-to-face interview	Mexico
Tidermark et al. [65]	110	80.0 (70–96)	79%	People with acute displaced femoral neck fractures; no severe cognitive impairment	Fractures ≤24 h. old	51% total hip replacement, 49% internal fixation	Structured interview at initial hospital stay and at 4 months	Sweden
Tidermark & Bergström [66]	59	82.9 (5.4) (70–92)	100%	Women with acute femoral neck fractures; no severe cognitive impairment	Fractures ≤24 h. old	N.A	Structured interview at initial hospital stay and at 6 months	Sweden
van Leeuwen et al. [67]	190	82.4 (7.7)	71.6%	Frail older adults living at home; recruited from general practices	N.A	N.A	Interview at participants' homes	Netherlands

Table 2 (continued)

Ref	Population			Disease/population characteristics			Instrumental administration	
	N*	Age Mean (SD, range)	% female	Disease/other characteristics of the study population/recruited from	Disease duration	Disease severity	Interview administration mode	Country
Walters et al. [68]	233	75²	66.5%	People with venous leg ulcers	present for ≥ 3 months	N.R	Community clinic setting	UK
EQ-5D-5L								
Aguirre et al. [69]	272	82.6 (8.1) (52–100)	61%	PwD; recruited from community (58%) and care homes (42%)	N.R	Mild to moderate	N.R	UK
Bjerk et al. [70]	155	82.7 (6.7)	79.3%	Older home-care recipients; experienced ≥ 1 fall in previous 12 months, not cognitively impaired; recruited from lists of people receiving professional home care	N.A	N.A	Interviewed in participants' homes by trained research assistants	Norway
Cheng et al. [71]	80	75 (13.88) (30–95)	59%	People with venous leg ulcers; recruited from 2 communities, 1 specialist & 1 hospital outpatient wound clinic	0–369 months	Venous insufficiency (41%), reduced mobility (81%)	N.R	Australia
Easton et al. [72]	143	85.7 (8.8) (49–99)	72%	Residential care setting	N.A	45% mild or moderate cognitive impairment, 25% dementia	N.R	Australia
Griffiths et al. [73]	726 (377 completed self-report measures)	85.6 (7.64) (57–102)	73.8%	PwD, living in care homes	N.R.	N.R	N.R	UK
Martin et al. [74]	1004	85.5 (58–103)	73.2%	PwD, living in residential care	N.A	N.A	Recruited from care homes	UK
Nikolova et al. [75]	1038	N.R (75% 75–84, 25% 85+)	52.7%	Community-dwelling older people; recruited from general practices	N.A	20.2% fit, 51.4% pre-frail, 28.4% frail	Face-to-face interview	UK

Table 2 (continued)

Ref	Population			Disease/population characteristics			Instrumental administration	
	N*	Age Mean (SD, range)	% female	Disease/other characteristics of the study population/recruited from	Disease duration	Disease severity	Interview administration mode	Country
Ratcliffe et al. [76]	240 (EQ-5D self-completed: n=82)	88.6 (5.6)	74.2%	Patients with hip fracture, living in residential aged care; recruited from acute orthopedic wards	N.A	93% moderate/severe dementia	N.R	Australia
EQ-5D-3L & EQ-5D-5L								
Bhadhuri et al. [41]	224	77.5 (5.35)	43.8%	People with multimorbidity and polypharmacy participating in a structured medication review RCT	N.A	N.A	Telephone interview	Belgium, Ireland, Netherlands, Switzerland

N.R not reported, N.A. not applicable, ADL activities of daily living, BL baseline, CG caregiver, FU follow-up, GP general practice/practitioner, MMSE mini mental state examination, Pw(A)D people with (Alzheimer's) dementia, RCT randomized controlled trial

*Sample size may differ for specific analyses, [1]only results for age group ≥ 75 years (or ≥ 80 years [53]) were extracted and included in this review, [2]median, [3]inter-quartile range

were recruited from different settings, e.g., residential care homes, home-care registries, general practices, falls prevention clinics, or the general population.

Evidence synthesis (Measurement properties)

The summarized results are presented in Table 3 (EQ-5D-3L) and Table 4 (EQ-5D-5L).

Reliability

In total, five studies assessed the reliability of the EQ-5D-3L index, with three reporting sufficient [39, 58, 67] and two reporting insufficient reliability [42, 52]. In one of the two studies of insufficient reliability [42], the time interval between measurements (6 months) was inappropriate (doubtful methodological quality). However, for the other study with insufficient reliability [52], no possible explanation could be found (similar population and/or time interval like in other studies reporting sufficient reliability [39, 58]). Thus, the overall rating of reliability of the EQ-5D-3L was inconsistent. Very low-quality evidence regarding the reliability of the individual dimensions of the EQ-5D-3L was available from one study [39], which found insufficient reliability based on Kappa coefficients between 0.34 and 0.59.

No study regarding the reliability of the EQ-5D-5L could be identified.

Convergent validity

Overall, convergent validity for both EQ-5D versions was supported by multiple studies, with the majority of hypotheses being supported at moderate to high quality of evidence.

As hypothesized, strong correlations between the EQ-5D-3L index and other instruments of HrQoL (SF-12, SF-6D, SF-36, HUI3) were found [40, 58, 65, 67]. At least moderate correlations were found with instruments of QoL (ICECAP-O, OPQOL-Brief, ASCOT, AQOL, QWB, QoL-AD) [44, 49–51, 57, 58, 61, 67], activities of daily living (ADL) (Barthel, Katz, BADL) [54, 58, 62, 64, 67], or single-scale instruments of general health or QoL [39, 50, 55, 57, 58, 63, 67]. Moreover, at least weak correlations with instruments of instrumental activities of daily living (IADL) (e.g., Lawton-Brody, NOSGER) [44, 54, 58, 62, 64] and comorbidities [58, 64] were found in the majority of studies. Results were inconsistent regarding the convergent validity of the EQ-5D-3L index with measures of depression/anxiety, which were hypothesized to be at least weakly correlated [57, 58, 60, 62].

Similarly, the EQ-5D-5L index was strongly correlated with the SF-6D as measure of HrQoL [70, 75]. At least moderate associations were found with QoL instruments (DEMQOL, DEMQOL-U, QOL-AD, SPVU-5D) [69, 71–73] (with the exception of the QoL-AD-NH [74]), as well as with a single-scale instrument for general health (EQ-VAS) [71] or a measure of ADL (MBI) [72, 76]. Results were inconsistent for associations with measures of cognitive status (Hypothesis 9, Table 1) [72, 74, 76], where one study found a positive correlation, although an association in the opposite direction was hypothesized [72].

Several studies [39, 41, 43, 44, 50, 51, 55, 56, 62–64, 68, 70–72, 75] also assessed convergent validity by correlating the EQ-5D index with the individual dimensions of the comparator instrument, the EQ-5D dimensions with a comparator instrument's summary score, or the EQ-5D dimensions with the comparator's dimensions (Tables S3 & S4, ESM 1). For both EQ-5D versions, the majority of results were in accordance with the hypotheses, thus, supporting the overall rating of convergent validity as sufficient.

Known-groups validity

Twelve studies assessed known-groups validity of the EQ-5D-3L index in a variety of populations [39, 42, 43, 47, 49, 51–55, 57, 68]. Overall, known-groups validity was inconsistent as < 75% of the results (67%) were in accordance with the hypotheses.

For the EQ-5D-5L index, known-groups validity was assessed in three studies [71, 72, 76]. The overall result was rated sufficient (78% of the hypotheses supported) and the quality of evidence was rated high.

Detailed information about the groups that the EQ-5D-3L and EQ-5D-5L were able to discriminate between can be found in Tables 3 & 4.

Responsiveness

Eight studies assessed responsiveness of the EQ-5D-3L index by examining the associations of change scores with other instruments [48, 49, 54, 56, 59, 65–67]. With one exception (AQoL) [49], the correlations with changes in instruments of HrQoL (SF-36, SF-12, NHP, 15D) [48, 65–67], QoL (ICECAP-O, ASCOT) [67], single-scale instruments of general health or QoL [67], ADL (Barthel, Katz) [54, 67], and IADL (NOSGER) [54] were weaker than hypothesized. Thus, responsiveness based on the comparison with other instruments was rated insufficient, and the summarized quality of evidence was rated high.

Ten studies assessed responsiveness of the EQ-5D-3L index based on comparisons between subgroups [41, 43, 45, 46, 54, 59, 61, 65, 66, 68]. These studies were primarily conducted on specific patient populations and assessed, e.g., the ability of the EQ-5D to differentiate between different outcomes after fractures or venous leg ulcers. Overall,

Table 3 Summary of findings—EQ-5D-3L

Measurement property	Summary	Overall rating	Quality of evidence
Reliability	Sub-dimensions [39]: Kappa: 0.34–0.59 ($n=45$)	–	**very low**
	Index: $ICC=0.58$–0.79 [39, 52, 58, 67], $r=0.67$ [42] ($n=439$)	±	**N/A**
Construct validity			
Convergent validity		**+ (91%)**	**high**
HrQoL instruments (*Hypothesis*: $r \geq 0.5$)	SF-6D [40], SF-36 [65], HUI3 [58], 15D [48]: **0.44** [48]–0.74; SF-12 MCS [67]: 0.36[a] ; SF-12 PCS [67]: 0.60 ($n=633$ or higher (n.c.r. [58]))	+(83%)	high
QoL instruments (*Hypothesis*: $r \geq 0.3$)	ICECAP-O [44, 61, 67], OPQOL-Brief [51], ASCOT [51, 67], AQoL [49], QWB scale [58], QoL-AD [50, 57]: 0.34–0.73 ($n≈1,588$ (n.c.r. [50, 57, 58, 61]))	+(100%)	high
General health/QoL (single-scale) (*Hypothesis*: $r \geq 0.3$)	Health GRS [67], EQ-VAS [39, 50, 58, 63], QoL GRS [67], Cantril's Self-Anchoring ladder [55], SF-36 general health [58], other [57] (3-pt ordinal scale): 0.34–0.52 ($n≈27,978$ (n.c.r. [50]))	+(100%)	high
ADL (*Hypothesis*: $r \geq 0.3$)	Barthel [54, 62], Katz [58, 64, 67]: **0.25** [58]–0.71; Bristol Activities of Daily Living Scale [60]: $\beta=-0.257$ ($n=1356$ or higher (n.c.r. [58]))	+(86%)	moderate
IADL (*Hypothesis*: $r \geq 0.1$)	Lawton & Brody [44, 58, 62], other [64], NOSGER [54]: **0.03** [44], 0.22–0.62 ($n=904$ or higher (n.c.r [58]))	+(80%)	moderate
Comorbidities (*Hypothesis*: $r \geq 0.1$)	Charlson [64], other [58] (0, 1, ≥ 2): 0.30–0.36 ($n=102$ or higher (n.c.r. [58]))	+(100%)	high
Cognitive status/dementia severity (*Hypothesis*: $r < 0.3$)	MMSE [44, 50, 54, 58, 64]: 0.07–0.20 ($n≈1,000$ (n.c.r. [50, 58]))	+(100%)	moderate
Depression/anxiety (*Hypothesis*: $r \geq 0.1$)	GDS [57, 58, 62]: **0.042** [62], 0.21–0.55; CSDD [60]: $\boldsymbol{\beta=-0.065}$ ($p > 0.05$); RAID [60]: $\beta=-0.168$ ($n≈1,280$ (n.c.r. [57, 58]))	±(60%)	N/A
Other instruments	$n≈770$ (n.c.r. [61])	+(100%)	high
(*Hypothesis*: $r \geq 0.3$)	OHS [61]: 0.70–0.77		
(*Hypothesis*: $r \geq 0.1$)	Pearlin Mastery Scale [67], Tinetti [62], VAS Pain [62]: 0.17–0.33		
(*Hypothesis*: $r < 0.3$)	CCCQ [67], PPA [44], SPPB [44]: 0.01–0.06		
Known-groups validity	$n≈31,176$ (n.c.r. [49, 53, 54, 57])	**± (67%)**	**N/A**
Supported for groups of…	Age[b] [42, 43, 49, 55, 68], sex [49, 51], social class[b] [49], education level[b] [51, 55], general health [51, 52, 57], mental & physical functioning (SF-12) [52], QoL-AD Score/Whole/Memory [52], IADL impairment (Lawton & Brody) [52, 57], disability severity [42], walking ability [68], number of medications [49], lower urinary tract symptom severity [53], obstructive airways disease (y/n) [47], depression (GDS) [52, 57], hospital stay (y/n) [42], multimorbidity [55], longstanding illness (y/n) [42], cognition (MMSE)[b] [52], confusion (mental test score)[b] [49], memory problems (GDS Memory)[b] [52],		
Rejected for groups of…	Age[b] [51], sex [55], living situation (alone vs. not alone/other arrangement) [49, 51, 55], informal care support (y/n) [51], marital status [55], GP visit (y/n) [42], outpatient attendance (y/n) [42], accident/emergency department attendance (y/n) [42], ADL impairment (higher vs. lower, Lawton-Brody) [52], only dementia vs. dementia + additional comorbidity [54], leg ulcer size and duration [68], functional impairment due to dementia [57], QoL-AD Life [52]		
Responsiveness			
Construct approach		**− (22%)**	**high**
HrQoL instruments (*Hypothesis*: $r \geq 0.5$)	SF-36 [65], NHP [66], SF-12 PCS [67], 15D[c] [48]: **0.23–0.39;** SF-12 MCS[a] [67]: **0.02** ($n=430$)	− (0%)	high
QoL instruments (*Hypothesis*: $r \geq 0.3$)	ICECAP-O [67], ASCOT [67]: **0.01–0.09;** AQoL [49]: 0.48 ($n≈219$ (n.c.r. [49]))	±(33%)	high
General health/QoL (single-scale) (*Hypothesis*: $r \geq 0.3$)	Health GRS [67], QoL GRS [67]: 0.12–0.14 ($n=149$)	− (0%)	high

Table 3 (continued)

Measurement property	Summary	Overall rating	Quality of evidence
ADL (*Hypothesis*: $r \geq 0.3$)	Barthel [54], Katz [67]: **0.04–0.19** (*n=484*)	− (0%)	moderate
IADL (*Hypothesis*: $r \geq 0.1$)	NOSGER [54]: **0.01** (*n=336*)	− (0%)	high
Cognitive status/dementia severity (*Hypothesis*: $r < 0.3$)	MMSE [54]: 0.00 (*n=369*)	+ (100%)	low
Other instruments	*n=371*	± (50%)	N/A
(*Hypothesis*: $r \geq 0.3$)	DASH [59]: 0.47		
(*Hypothesis*: $r \geq 0.1$)	Pearlin Mastery Scale [67], Activity inventory [56]: **0.02–0.06**		
(*Hypothesis*: $r < 0.3$)	CCCQ [67]: 0.09		
Comparison between subgroups	*n≈1,711* (*n.c.r.* [54])	**+ (79%)**	**moderate**
Supported for groups of…	Improvement/worsening on the Barthel index [41], knee replacement vs. femur fracture [43], femur fracture vs. stroke [43], fallers vs. non-fallers [45], complication vs. non-complication after femoral neck fracture [46], deterioration in health status (CGI-I) [54], less good vs. good outcome after femoral neck fracture (pain and/or needing walking aids) [65], perceived health change and healing status in people with venous leg ulcers [68], complications/non-complications after femoral neck fracture [46], improvement/deterioration status (DASH) after proximal humeral fracture [59], death/non-death after hip fracture [61], displaced/undisplaced femoral neck fractures [66]		
Rejected for groups of…	Improvement/worsening on the EQ-VAS [41], hip replacement vs. femur fracture [43]; healed vs. non-healed leg ulcers at 3 months follow-up [68], revision after hip fracture [61]		
Before and after intervention			
Supported for…	Deterioration/improvement of HrQoL over time after hip or proximal humeral fracture [59, 61] (*n=340*)	+ (100%)	high
Rejected for …	Low-vision rehabilitation [56] (*n=77*)	− (0%)	moderate

Unless otherwise indicated, reported numbers refer to absolute correlation coefficients, correlation coefficients printed in **bold** indicate results for which the hypotheses were rejected

y/n yes/no, *n.s.* not significant, *N/A* not applicable, *N/R* not reported, *r* correlation coefficient, *β* regression coefficient, *n* sample size, *n.c.r.* not clearly reported, *ADL* activities of daily living, *ASCOT* adult social care outcomes toolkit, *AQoL* assessment of quality of life, *CCCQ* client-centered care questionnaire, *CGI-I* clinical global impression of improvement, *CSDD* Cornell Scale for depression in dementia, *DASH* disabilities of arm, shoulder, and hand, *EQ-VAS* Visual Analogue Scale, *GDS* Geriatric Depression Scale, *GRS* Global Rating Scale, *HrQoL* health-related quality of life, *HUI3* Health Utilities Index, *IADL* instrumental activities of daily living, *ICC* intraclass correlation coefficient, *ICECAP-O* ICEpop CAPability measure for older people, *MCS* mental health component summary, *MMSE* mini-mental state examination, *NHP* Nottingham Health Profile, *NOSGER* Nurses' Observation Scale for Geriatric Patients, *OHS* oxford hip score, *OPQOL-Brief* older people's quality-of-life brief questionnaire, *PCS* physical health component summary, *PPA* physiological profile assessment, *QoL* quality of life, *QoL-AD* qualityof life in Alzheimer's diseases, *QoL GRS* Quality-of-Life Global Rating Scale, *QWB* quality of well-being, *RAID* Rating of Anxiety in Dementia Scale, *SF-36* 36-item short-form health survey, *SF-12* 12-item short-form health survey, *SF-6D* six-dimensional short form, *SPPB* short physical performance battery, *VAS Pain* visual analogue scale for pain

[a]deviating hypothesis: $r \geq 0.1$

[b]no relevant difference between groups hypothesized

[c]no calculation of correlation, instead comparison of EQ-5D & 15D in terms of proportions of changes stratified according to the minimally important difference values

Table 4 Summary of findings—EQ-5D-5L

Measurement property	Summary or pooled results	Overall rating	Quality of evidence
Reliability	N/R		
Construct validity			
Convergent validity		**+ (84%)**	**High**
HrQoL instruments (*Hypothesis*: $r \geq 0.5$)	SF-6D: 0.71 [70], $ICC = 0.61$ [75] ($n \approx 1193$ (*n.c.r.* [75]))	+ (100%)	High
QoL instruments (*Hypothesis*: $r \geq 0.3$)	DEMQOL [69], DEMQOL-U [72], QOL-AD [69, 73]: 0.30–0.48 QOL-AD-NH [74]: **0.28;** SPVU-5D [71]: $ICC = 0.55$ ($n \approx 1417$ (*n.c.r.* [71]))	+ (83%)	High
General health/QoL (single scale) (*Hypothesis*: $r \geq 0.3$)	EQ-VAS [71]: 0.39 ($n \approx 75$ (*n.c.r.*))	+ (100%)	Moderate
ADL (*Hypothesis*: $r \geq 0.3$)	MBI [72, 76]: 0.46–0.49 ($n = 225$)	+ (100%)	High
Cognitive status (*Hypothesis*: $r < 0.3$)	Pas-Cog* [72]: **0.24;** MMSE [76]: 0.22; CDR [74]: 0.025 ($n = 1116$)	± (67%)	N/A
Other instruments	$n = 1113$	+ (80%)	High
(*Hypothesis*: $r \geq 0.1$)	CSDD [76], PainAd [76]: 0.33–0.45; FAST [74]: **0.049**		
(*Hypothesis*: $r < 0.3$)	CMAI [74], NPI-Q [72]: 0.1		
Known-groups validity	$n \approx 306$ (*n.c.r.* [71])	**+ (78%)**	**High**
Supported for Groups of…	Age[a] [71], general health (EQ-VAS) [71], leg ulcer healing status [71], physical functioning/ADL (MBI) [72, 76], pain (PainAd) [76], depression (CSSD) [76]		
Rejected for groups of…	cognitive impairment (PAS-Cog)* [72], ulcer duration [71]		
Responsiveness			
Construct approach		**+ (75%)**	**High**
QoL instruments (*Hypothesis*: $r \geq 0.3$)	QOL-AD-NH [74]: $\boldsymbol{\beta \approx 0.007}$ ($p < 0.05$) ($n \approx 261$(*n.c.r.*))	− (0%)	Moderate
Cognitive status (*Hypothesis*: $r < 0.3$)	CDR [74]: $\beta =$ n.s ($n \approx 261$(*n.c.r.*))	+ (100%)	High
Other instruments	$n \approx 396$ (*n.c.r.* [74])	+ (83%)	High
(*Hypothesis*: $r \geq 0.3$)	BBS [70]: Elasticity = 0.54		
(*Hypothesis*: $r \geq 0.1$)	30 s STS [70], 4 m walk test [70], FES-I [70]: Elasticity = 0.09–0.24; FAST [74]: $\boldsymbol{\beta = \text{n.s}}$		
(*Hypothesis*: $r < 0.3$)	CMAI [74]: $\beta =$ n.s		
Comparison between subgroups	$n = 269$	**+ (75%)**	**High**
Supported for groups of…	Improvement/worsening on the Barthel index [41], healing status and duration of venous leg ulcers [71]		
Rejected for groups of…	Improvement/worsening on the EQ-VAS [41]		
Before and after intervention	N/R		

+ sufficient, − insufficient, ± inconsistent, *y/n* yes/no, *n.s.* not significant, *r* correlation coefficient, *β* regression coefficient, *n* sample size, *n.c.r.* not clearly reported, *N/R* not reported, *BBS* Berg Balance Scale, *CDR* clinical dementia rating, *CMAI* Cohen-Mansfield Agitation Inventory, *CSDD* Cornell Scale for Depression in Dementia, *DEMQOL* dementia quality of life, *EQ-VAS* Visual Analog Scale, *FAST* functional assessment staging, *FES-I* Falls Efficacy Scale International, *ICC* intraclass correlation coefficient, *MBI* Modified Barthel Index, *MMSE* mini-mental state examination, *PainAd* Pain Assessment in Advanced Dementia Scale, *PAS-Cog* Psychogeriatric Assessment Scales-Cognitive Impairment Scale, *QoL* quality of life, *QoL-AD* quality of life in Alzheimer's disease, *QOL-AD-NH* quality of life in Alzheimer's disease nursing home version, *SF-6D* six-dimensional short-form health survey, *30 s STS* 30-second sit-to-stand test, *SPVU-5D* five-dimensional sheffield-preference-based venous ulcer questionnaire

*result in the opposite of the hypothesized direction (H9)

[a]no relevant difference between groups hypothesized

moderate-quality evidence for sufficient responsiveness of the EQ-5D-3L based on comparisons between subgroups was found, as 79% of the hypotheses were supported.

Three studies [56, 59, 61] examined responsiveness by testing hypotheses regarding change in the EQ-5D-3L index in response to an intervention. Two hypotheses regarding the improvement or deterioration of HrQoL after fracture were supported, whereas, opposed to the hypothesis, low vision rehabilitation did not change HrQoL.

For the EQ-5D-5L index, two studies [70, 74] assessed responsiveness based on comparisons with other instruments. 75% of the results were in accordance with the

hypotheses and, thus, were rated as sufficient at high quality of evidence. The correlations of change scores were as high (or low) as hypothesized between the EQ-5D-5L and measures of cognitive status or agitation (CDR, CMAI) [74], measures of physical function (BBS, 30 s STS, 4 m walk test) [70] but were lower than hypothesized between the EQ-5D-5L and a QoL instrument (QOL-AD-NH) [74] or a measure of functional symptoms in dementia (FAST) [74].

Two studies examined responsiveness of the EQ-5D-5L index in terms of subgroup comparisons [41, 71]. 75% of the hypotheses were supported and, thus, the overall result was sufficient. The quality of evidence was rated high.

Results not included in the qualitative synthesis

Some results were not included in the qualitative synthesis as no specific results (e.g., correlation coefficients) were reported. Regarding convergent validity, Michalowsky et al. [57] found a poor association (not further specified) between the EQ-5D-3L index and IADL. Other authors examined the association between the EQ-5D dimensions with ADL and found significant associations between several dimensions but did not provide information about the strength of the association [39, 43]. Moreover, the authors assessed known-groups validity and found, e.g., that women were more anxious than men [39] and that people with disability had lower HrQoL than people with no disability [43]. However, it could not be evaluated whether the differences were clinically important because the mean EQ-5D of each group was not reported.

Discussion

The current study synthesized reliability, validity, and responsiveness of the EQ-5D in a population of middle-old and oldest-old people. Regarding reliability, results were inconsistent for the EQ-5D-3L, and for the EQ-5D-5L, studies were entirely lacking. This may pose a problem in contexts where the EQ-5D is used at different time points to quantify a 'true' difference or change in HrQoL, such as in economic evaluations. Previous reviews report mixed results on the reliability of the EQ-5D in people with dementia (moderate to strong) [16] and sufficient reliability in people with diabetes or stroke [77, 78]. Another review further suggests sufficient reliability of the EQ-5D-5L in various patient groups (e.g., osteoarthritis, diabetes and cancer patients, cardiovascular and liver diseases) and general population samples [79]. However, so far, the evidence on reliability for both the EQ-5D-3L and EQ-5D-5L is relatively limited and entirely lacking for certain patient groups.

For both EQ-5D versions, high-quality evidence of sufficient convergent validity was found. It should be noted

that high correlations with other generic instruments (e.g., SF-36/-12, SF-6D, HUI3) do not necessarily support the use of the EQ-5D in middle-old to oldest-old people, as it does not preclude that both instruments do not capture aspects that are important to the population of interest. In some cases, convergent validity was assessed by correlations with instruments which were collected only in a single, specific study (e.g., OHS, Pearlin Mastery Scale). These results summarized as "other instruments" despite measuring different constructs in Table 3 and 4, may not be generally relevant for the population aged 75+ but were mostly in accordance with the hypotheses.

Known-groups validity of the EQ-5D-3L was inconsistent. One potential explanation could be a ceiling effect of the EQ-5D-3L, which may have compromised its ability to discriminate between known groups. Moreover, it can be questioned whether the groups for evaluating known-groups validity are relevant (e.g., marital status, living alone vs. not alone). Similarly, it could be questioned whether it is reasonable to examine, e.g., convergent validity of the EQ-5D with instruments measuring constructs which are hardly related to HrQoL (e.g., CCCQ, PPA, SPPB). The evaluation of measurement properties should be theory driven and not exploratory by using all available variables from studies that were initially designed for a different purpose. More precise preliminary hypotheses of associations between measures in studies analyzing an instrument's measurement properties would, therefore, be desirable. In addition, rather "soft" hypotheses regarding the strength of the association between two instruments were defined in this review, e.g., by not setting an upper limit for correlations between instruments measuring related but dissimilar constructs ($r \geq 0.3$) or weakly related constructs ($r \geq 0.1$). This was done to avoid "penalizing" relatively strong correlations between instruments that were assumed to be not necessarily but potentially highly correlated (e.g., EQ-5D and ADL instruments). Since, according to the COSMIN methodology, the synthesized evaluation of a measurement property is based on a majority principle ($\geq 75\%$ of the hypotheses supported), these aspects could have influenced the (synthesized) results. For the EQ-5D-5L, high-quality evidence of sufficient known-groups validity was found. There, the selection of groups that the EQ-5D was expected to differentiate between seemed to be less arbitrary, but overall, the results were based on only three studies. The COSMIN methodology recommends judging an instrument's ability to discriminate between relevant groups based on clinically important rather than statistically significant differences [34]. While being aware that there is no single MCID for EQ-5D index values since it varies by population characteristics [80], in the absence of specific MCIDs for each country-specific tariff and disease group of the individual studies included in this review, MCIDs commonly used in

previous literature were nevertheless used but could have influenced the results regarding known-groups validity.

Responsiveness was insufficient (high-quality evidence) for the EQ-5D-3L when correlated with instruments being hypothesized to be related (e.g., other (Hr)QoL instruments). However, it seemed to be responsive to outcomes after fracture or healing status of leg ulcers [43, 46, 59, 61, 65, 66, 68]. These are conditions with substantial changes in health, where the EQ-5D has previously been shown to be more likely to be responsive (in an older population) [12, 18]. Although responsiveness of the EQ-5D-5L (construct approach) was found sufficient according to the majority principle of the COSMIN methodology, the evidence was limited as it was based on only two studies which used very study-specific instruments to evaluate responsiveness (e.g., 30 s STS) [70, 74]. These instruments were hypothesized to be only weakly associated with the EQ-5D and were, therefore, not responsive to changes in HrQoL.

Overall, the results regarding the responsiveness of the EQ-5D suggest that at least the EQ-5D-3L is hardly able to adequately reflect clinical changes over time. In turn, clinically relevant changes may remain undetected; thus, intervention effects may be underestimated based on the EQ-5D. For example, economic evaluations of fall prevention programs showed that clinical effects could not be found on HrQoL [81–83]. This does not seem to be an exclusive problem of the EQ-5D but also of other generic HrQoL instruments, such as the SF-36 or SF-12 [82, 83]. So far, the evidence on responsiveness of the EQ-5D is mainly based on studies using the EQ-5D-3L. The sparse evidence on the responsiveness of the EQ-5D-5L is not limited to the population of middle old to oldest old but is also found in general for other populations [79]. Moreover, the majority of the included studies reported substantial ceiling effects, which may limit the ability to capture small changes at the upper end of HrQoL. Ceiling effects were found to be particularly common among people with dementia [15], who make up a large proportion in the current study. Generally, the EQ-5D-5L was found to reduce this ceiling effect [84, 85]. However, it persists in general population studies but also in some patient populations [79]. Further studies are needed, which evaluate the responsiveness of the EQ-5D-5L to change in, e.g., other (age or disease specific) (Hr)QoL instruments. It would be of particular interest to examine whether the EQ-5D-5L is more responsive than the EQ-5D-3L which was insufficiently responsive in this respect.

The approach to primarily focus on HrQoL in the form of health utility gains in economic evaluations has been criticized for excluding aspects of QoL beyond health [23, 86]. Furthermore, HrQoL instruments such as the EQ-5D or the SF-12/SF-36 are mainly functioning oriented and, thus, do not reflect the breadth of the concept of health as stated in the WHO definition [21], e.g., social aspects of health fall short or are not assessed differentiated enough. This seems to be especially relevant to older people as it was found that not only health but also social domains are important to their overall QoL [23, 87]. Therefore, other instruments were and are currently being developed, which may provide an alternative or complement to measure (Hr)QoL based on a broader or more comprehensive framework of health or well-being in the future. Some age- or disease-specific QoL instruments exist, and the current study showed that although being moderately to strongly associated with the EQ-5D when assessed at a single time point (sufficient convergent validity), changes on these instruments are not reflected on the EQ-5D (insufficient responsiveness). This suggests that the EQ-5D is not able to capture changes in (Hr)QoL that are important to older people. However, the existing age- or disease-specific instruments differ in domains of (Hr)QoL that are captured [6] and, thus, pose a problem for the comparability of intervention effects across diseases and populations. Moreover, the lack of preference-based value sets for some of these instruments (e.g., for the WHOQOL-OLD, an older people-specific QoL instrument [87]) or value sets being only available for the population in the country where the instruments were developed, impedes their use in economic evaluations. Another recently developed instrument is the PROMIS-29, a health profile measure from the Patient-Reported Outcomes Measurement Information System® (PROMIS®) [88–90] that captures health in a broader sense than the EQ-5D. Although value sets are available for the PROMIS-29 [89–92], they are so far only available for the US. Moreover, the 'Extending the QALY' research project is currently developing the EQ-HWB, a broad measure of QoL for use in economic evaluations across health and social care (https://scharr.dept.shef.ac.uk/e-qaly/), and thus, could be a potential alternative to the EQ-5D in the future. However, these age-unspecific instruments carry the risk that scoring algorithms used to derive the utility index are based on the preferences of the general adult populations, whose preferences for health may differ from those of older people [6, 24]. Another research group is seeking to address this issue and is currently developing an instrument for quality assessment and economic evaluation that adequately captures the aspects of quality of life that are important to older people, using a person-centered approach [93, 94]. Consequently, as long as there is no single preference-based generic instrument that comprehensively captures relevant aspects of (Hr)QoL in middle-old and oldest-old people or its use is limited in certain situations (e.g., lack of country/population-specific tariffs), age- or disease-specific instruments should be used as complement to the EQ-5D and help interpreting the results of (cost-)effectiveness analyses (e.g., whether the effects of an intervention are likely to be underestimated).

Beyond these alternative instruments, several "bolt-on" dimensions to the EQ-5D have been proposed and a wide variety of methods have been applied to identify or select relevant bolt-on dimensions [95]. Finch, Brazier, Mukuria, and Bjorner [96] identified hearing, sleep, cognition, energy, and relationships as potentially relevant bolt-on dimensions, and some studies have shown that higher severity levels in the bolt-on dimensions impact the health state values or preferences for the health state [97–99]. Recently, Chen and Olsen [100] proposed vitality, sleep, social relationships, and community connectedness as bolt-on dimensions. They argue that adding these four dimensions would provide a solution to assess HrQoL in a single, brief instrument, but still include all key dimensions of the conceptual map of HrQoL by Olsen and Misajon, [21] and, thus, capture health and well-being more broadly than current EQ-5D instruments. However, to use the additional information from the bolt-on dimensions in economic evaluations, the bolt-on dimension scores would need to be incorporated into the utility index, which would require new valuation studies. Moreover, extensive testing on whether the bolt-on dimensions improve psychometric performance of the EQ-5D would be needed, in general, but also particularly in middle-old and oldest-old people.

A large number of the included studies ($n = 13$) assessed the measurement properties of the EQ-5D in people with dementia or cognitive impairment. As part of the validation, the association between (change in) cognitive status and (change in) the EQ-5D was examined [44, 49, 52, 54, 58, 64, 72, 74, 76]. However, the relationship between cognition and (Hr)QoL seems to be complex [101, 102], which made it difficult to formulate (generic) hypotheses regarding the direction and strength of the association in this study.

This review deliberately did not focus on the comparison of self- and proxy-rated EQ-5D scores and did not consider correlations between the self-rated EQ-5D and proxy-rated other (Hr)QoL instruments in the synthesis. (Hr)QoL is a subjective concept; therefore, it is not surprising that different people evaluate it differently, especially when self-perception is impaired by a condition such as dementia, where proxies typically rate the HrQoL of a person with dementia lower than the person him/herself [15, 16]. It is not possible to determine whose rating is more "correct." However, it is important to be aware of these variations and to select the administration mode depending on the perspective from which the benefits of an intervention are to be measured.

This study applied the updated COSMIN methodology to systematically review the measurement properties of the EQ-5D in a middle-old and oldest-old population. However, several limitations must be acknowledged. First, only studies which directly aimed to examine the measurement properties of the EQ-5D were included, whereas studies providing indirect evidence on measurement properties (e.g., by correlating the EQ-5D with instruments being hypothetically related) were not included. Second, the generalizability of the results may be limited: although this study was deliberately not restricted to specific populations such as disease groups, it is not clear, whether the results apply to the general population of middle-old to oldest-old adults as, e.g., a large share of the included studies included only people with dementia. Moreover, the results do not exclusively apply to the population aged 75+ as a number of persons < 75 years are also included in some of the studies. To date, there have been few studies focusing exclusively on the population aged 75 years and older, representing a gap in research. Such studies could allow a comparison between the measurement properties of the EQ-5D between younger-old (e.g., aged 60+) and middle-old to oldest-old people, which was not directly possible based on the current data. Finally, the evidence stems exclusively from western, industrialized countries and, therefore, may not be transferable to other countries or regions.

Conclusion

The results of this systematic review are relevant as improving the care and maintaining the health and QoL of an older population is a political goal in many countries. Thereby, the results may be of interest to decision makers, but also to researchers planning, designing, or evaluating interventions for older people.

Based on the findings of this study, both EQ-5D versions seem to have sufficient convergent validity and may, therefore, be used in cross-sectional studies to assess HrQoL. However, caution is advised when using the EQ-5D to assess change in HrQoL, as the EQ-5D-3L was found to be insufficiently responsive to change (except for conditions with substantial changes in health) and results regarding the reliability were inconsistent. As specifically for the EQ-5D-5L little evidence on reliability and responsiveness is available so far, further research might be needed in this regard. If responsiveness cannot be demonstrated, either using additional disease- or age-specific instruments or considering the use of an alternative, more comprehensive instrument of (Hr)QoL might be advisable, especially for economic evaluations. Promising research is currently underway to develop new, more comprehensive instruments that will better capture the aspects of QoL that are important to older people. However, there is still a long way to go to verify their measurement properties, generate population- and country-specific value sets, and thus, be broadly applicable to economic evaluations.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

1. United Nations, Department of Economic and Social Affairs, & Population Division. (2019). *World population prospects 2019* (Vol. 2). Demographic Profiles.
2. Janssen, M. F., Szende, A., Cabases, J., Ramos-Goñi, J. M., Vilagut, G., & König, H. H. (2019). Population norms for the EQ-5D-3L: A cross-country analysis of population surveys for 20 countries. *The European Journal of Health Economics, 20*(2), 205–216. https://doi.org/10.1007/s10198-018-0955-5
3. Marten, O., & Greiner, W. (2021). EQ-5D-5L reference values for the German general elderly population. *Health and Quality of Life Outcomes, 19*(1), 76. https://doi.org/10.1186/s12955-021-01719-7
4. EuroQol Group. (1990). EuroQol–a new facility for the measurement of health-related quality of life. *Health Policy, 16*(3), 199–208. https://doi.org/10.1016/0168-8510(90)90421-9
5. Kennedy-Martin, M., Slaap, B., Herdman, M., van Reenen, M., Kennedy-Martin, T., Greiner, W., Busschbach, J., & Boye, K. S. (2020). Which multi-attribute utility instruments are recommended for use in cost-utility analysis? A review of national health technology assessment (HTA) guidelines. *The European Journal of Health Economics, 21*(8), 1245–1257. https://doi.org/10.1007/s10198-020-01195-8
6. Cleland, J., Hutchinson, C., Khadka, J., Milte, R., & Ratcliffe, J. (2019). A review of the development and application of generic preference-based instruments with the older population. *Applied Health Economics and Health Policy, 17*(6), 781–801. https://doi.org/10.1007/s40258-019-00512-4
7. National Institute for Health and Care Excellence. (2013). Guide to the methods of technology appraisal 2013 [Internet]. *Process and Methods Guides No9.* National Institute for Health and Care Excellence.
8. Brazier, J., Connell, J., Papaioannou, D., Mukuria, C., Mulhern, B., Peasgood, T., Jones, M. L., Paisley, S., O'Cathain, A., Barkham, M., Knapp, M., Byford, S., Gilbody, S., & Parry, G. (2014). A systematic review, psychometric analysis and qualitative assessment of generic preference-based measures of health in mental health populations and the estimation of mapping functions from widely used specific measures. *Health Technology Assessment, 18*(34), 1–188. https://doi.org/10.3310/hta18340
9. Mulhern, B., Mukuria, C., Barkham, M., Knapp, M., Byford, S., Soeteman, D. R., & Brazier, J. (2014). Using generic preference-based measures in mental health: Psychometric validity of the EQ-5D and SF-6D. *British Journal of Psychiatry, 205*(3), 236–243. https://doi.org/10.1192/bjp.bp.112.122283
10. Davis, S., & Wailoo, A. (2013). A review of the psychometric performance of the EQ-5D in people with urinary incontinence. *Health and Quality of Life Outcomes, 11*(1), 20. https://doi.org/10.1186/1477-7525-11-20
11. Yang, Y., Brazier, J., & Longworth, L. (2015). EQ-5D in skin conditions: An assessment of validity and responsiveness. *The European Journal of Health Economics, 16*(9), 927–939. https://doi.org/10.1007/s10198-014-0638-9
12. Haywood, K. L., Garratt, A. M., & Fitzpatrick, R. (2005). Quality of life in older people: A structured review of generic self-assessed health instruments. *Quality of Life Research, 14*(7), 1651–1668. https://doi.org/10.1007/s11136-005-1743-0
13. Kuspinar, A., & Mayo, N. E. (2014). A review of the psychometric properties of generic utility measures in multiple sclerosis. *PharmacoEconomics, 32*(8), 759–773. https://doi.org/10.1007/s40273-014-0167-5
14. Tordrup, D., Mossman, J., & Kanavos, P. (2014). Responsiveness of the EQ-5D to clinical change: Is the patient experience adequately represented? *International Journal of Technology Assessment in Health Care, 30*(1), 10–19. https://doi.org/10.1017/s0266462313000640
15. Hounsome, N., Orrell, M., & Edwards, R. T. (2011). EQ-5D as a quality of life measure in people with dementia and their carers: Evidence and key issues. *Value in Health, 14*(2), 390–399. https://doi.org/10.1016/j.jval.2010.08.002
16. Li, L., Nguyen, K. H., Comans, T., & Scuffham, P. (2018). Utility-based instruments for people with dementia: A systematic review and meta-regression analysis. *Value in Health, 21*(4), 471–481. https://doi.org/10.1016/j.jval.2017.09.005
17. Finch, A. P., Brazier, J. E., & Mukuria, C. (2018). What is the evidence for the performance of generic preference-based measures? A systematic overview of reviews. *The European Journal of Health Economics, 19*(4), 557–570. https://doi.org/10.1007/s10198-017-0902-x

18. Payakachat, N., Ali, M. M., & Tilford, J. M. (2015). Can the EQ-5D detect meaningful change? A systematic review. *PharmacoEconomics, 33*(11), 1137–1154. https://doi.org/10.1007/s40273-015-0295-6

19. Brazier, J., Roberts, J., Tsuchiya, A., & Busschbach, J. (2004). A comparison of the EQ-5D and SF-6D across seven patient groups. *Health Economics, 13*(9), 873–884. https://doi.org/10.1002/hec.866

20. Janssen, M. F., Pickard, A. S., Golicki, D., Gudex, C., Niewada, M., Scalone, L., Swinburn, P., & Busschbach, J. (2013). Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: A multi-country study. *Quality of Life Research, 22*(7), 1717–1727. https://doi.org/10.1007/s11136-012-0322-4

21. Olsen, J. A., & Misajon, R. (2020). A conceptual map of health-related quality of life dimensions: Key lessons for a new instrument. *Quality of Life Research, 29*(3), 733–743. https://doi.org/10.1007/s11136-019-02341-3

22. Sutton, E. J., & Coast, J. (2014). Development of a supportive care measure for economic evaluation of end-of-life care using qualitative methods. *Palliative Medicine, 28*(2), 151–157. https://doi.org/10.1177/0269216313489368

23. Milte, C. M., Walker, R., Luszcz, M. A., Lancsar, E., Kaambwa, B., & Ratcliffe, J. (2014). How important is health status in defining quality of life for older people? An exploratory study of the views of older South Australians. *Applied Health Economics and Health Policy, 12*(1), 73–84. https://doi.org/10.1007/s40258-013-0068-3

24. Ratcliffe, J., Lancsar, E., Flint, T., Kaambwa, B., Walker, R., Lewin, G., Luszcz, M., & Cameron, I. D. (2017). Does one size fit all? Assessing the preferences of older and younger people for attributes of quality of life. *Quality of Life Research, 26*(2), 299–309. https://doi.org/10.1007/s11136-016-1391-6

25. Marten, O., Brand, L., & Greiner, W. (2021). Feasibility of the EQ-5D in the elderly population: A systematic review of the literature. *Quality of Life Research*. https://doi.org/10.1007/s11136-021-03007-9

26. Bulamu, N. B., Kaambwa, B., & Ratcliffe, J. (2015). A systematic review of instruments for measuring outcomes in economic evaluation within aged care. *Health and Quality of Life Outcomes, 13*(1), 179. https://doi.org/10.1186/s12955-015-0372-8

27. Makai, P., Brouwer, W. B. F., Koopmanschap, M. A., Stolk, E. A., & Nieboer, A. P. (2014). Quality of life instruments for economic evaluations in health and social care for older people: A systematic review. *Social Science and Medicine, 102*, 83–93. https://doi.org/10.1016/j.socscimed.2013.11.050

28. Prinsen, C. A. C., Mokkink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., de Vet, H. C. W., & Terwee, C. B. (2018). COSMIN guideline for systematic reviews of patient-reported outcome measures. *Quality of Life Research, 27*(5), 1147–1157. https://doi.org/10.1007/s11136-018-1798-3

29. Gottschalk, S., König, H. H., Nejad, M., & Dams, J. (2020). Psychometric properties of the EQ-5D for the assessment of health-related quality of life in the population of middle-old and oldest-old persons: Study protocol for a systematic review. *Frontiers in Public Health, 8*, 578073. https://doi.org/10.3389/fpubh.2020.578073

30. Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The, P. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine, 6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

31. Terwee, C. B., Jansma, E. P., Riphagen, I. I., & de Vet, H. C. W. (2009). Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. *Quality of Life Research, 18*(8), 1115–1123. https://doi.org/10.1007/s11136-009-9528-5

32. Mokkink, L. B., de Vet, H. C. W., Prinsen, C. A. C., Patrick, D. L., Alonso, J., Bouter, L. M., & Terwee, C. B. (2018). COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. *Quality of Life Research, 27*(5), 1171–1179. https://doi.org/10.1007/s11136-017-1765-4

33. Terwee, C. B., Bot, S. D., de Boer, M. R., van der Windt, D. A., Knol, D. L., Dekker, J., Bouter, L. M., & de Vet, H. C. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of clinical epidemiology, 60*(1), 34–42. https://doi.org/10.1016/j.jclinepi.2006.03.012

34. Mokkink, L. B., Prinsen, C. A. C., Patrick, D. L., Alonso, J., Bouter, L. M., de Vet, H. C., & Terwee, C. B. (2018). COSMIN methodology for systematic reviews of patient-reported outcome measures. *Quality of Life Research, 27*(5), 1147–1157.

35. Walters, S. J., & Brazier, J. E. (2005). Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. *Quality of Life Research, 14*(6), 1523–1532. https://doi.org/10.1007/s11136-004-7713-0

36. McClure, N. S., Sayah, F. A., Xie, F., Luo, N., & Johnson, J. A. (2017). Instrument-defined estimates of the minimally important difference for EQ-5D-5L index scores. *Value in Health, 20*(4), 644–650. https://doi.org/10.1016/j.jval.2016.11.015

37. Prinsen, C. A. C., Vohra, S., Rose, M. R., Boers, M., Tugwell, P., Clarke, M., Williamson, P. R., & Terwee, C. B. (2016). How to select outcome measurement instruments for outcomes included in a "Core Outcome Set"—a practical guideline. *Trials, 17*(1), 449. https://doi.org/10.1186/s13063-016-1555-2

38. GRADE Handbook (2013). Handbook for grading the quality of evidence and the strength of recommendations using the GRADE approach. Retrieved February18, 2021, from https://gdt.gradepro.org/app/handbook/handbook.html

39. Ankri, J., Beaufils, B., Novella, J. L., Morrone, I., Guillemin, F., Jolly, D., Ploton, L., & Blanchard, F. (2003). Use of the EQ-5D among patients suffering from dementia. *Journal of clinical epidemiology, 56*(11), 1055–1063. https://doi.org/10.1016/s0895-4356(03)00175-6

40. Barton, G. R., Sach, T. H., Avery, A. J., Jenkinson, C., Doherty, M., Whynes, D. K., & Muir, K. R. (2008). A comparison of the performance of the EQ-5D and SF-6D for individuals aged > or = 45 years. *Health Economics, 17*(7), 815–832. https://doi.org/10.1002/hec.1298

41. Bhadhuri, A., Kind, P., Salari, P., Jungo, K. T., Boland, B., Byrne, S., Hossmann, S., Dalleur, O., Knol, W., Moutzouri, E., O'Mahony, D., Murphy, K. D., Wisselink, L., Rodondi, N., & Schwenkglenks, M. (2020). Measurement properties of EQ-5D-3L and EQ-5D-5L in recording self-reported health status in older patients with substantial multimorbidity and polypharmacy. *Health and Quality of Life Outcomes, 18*(1), 317. https://doi.org/10.1186/s12955-020-01564-0

42. Brazier, J. E., Walters, S. J., Nicholl, J. P., & Kohler, B. (1996). Using the SF-36 and EuroQol on an elderly population. *Quality of Life Research, 5*(2), 195–204. https://doi.org/10.1007/bf00434741

43. Coast, J., Peters, T. J., Richards, S. H., & Gunnell, D. J. (1998). Use of the EuroQoL among elderly acute care patients. *Quality of Life Research, 7*(1), 1–10. https://doi.org/10.1023/a:1008857203434

44. Davis, J. C., Bryan, S., McLeod, R., Rogers, J., Khan, K., & Liu-Ambrose, T. (2012). Exploration of the association between quality of life, assessed by the EQ-5D and ICECAP-O, and falls risk, cognitive function and daily function, in older adults with mobility impairments. *BMC Geriatrics, 12*, 65. https://doi.org/10.1186/1471-2318-12-65

45. Davis, J. C., Best, J. R., Dian, L., Khan, K. M., Hsu, C. L., Chan, W., Cheung, W., & Liu-Ambrose, T. (2017). Are the EQ-5D-3L and the ICECAP-O responsive among older adults with impaired

mobility? Evidence from the Vancouver falls prevention cohort study. *Quality of Life Research, 26*(3), 737–747. https://doi.org/10.1007/s11136-016-1487-z

46. Frihagen, F., Grotle, M., Madsen, J. E., Wyller, T. B., Mowinckel, P., & Nordsletten, L. (2008). Outcome after femoral neck fractures: A comparison of Harris hip score, Eq-5d and Barthel index. *Injury, 39*(10), 1147–1156. https://doi.org/10.1016/j.injury.2008.03.027

47. Hazell, M., Frank, T., & Frank, P. (2003). Health related quality of life in individuals with asthma related symptoms. *Respiratory Medicine, 97*(11), 1211–1218. https://doi.org/10.1016/S0954-6111(03)00249-X

48. Heiskanen, J., Tolppanen, A.-M., Roine, R. P., Hartikainen, J., Hippeläinen, M., Miettinen, H., & Martikainen, J. (2016). Comparison of EQ-5D and 15D instruments for assessing the health-related quality of life in cardiac surgery patients. *European Heart Journal, 2*(3), 193–200. https://doi.org/10.1093/ehjqcco/qcw002

49. Holland, R., Smith, R. D., Harvey, I., Swift, L., & Lenaghan, E. (2004). Assessing quality of life in the elderly: A direct comparison of the EQ-5D and AQoL. *Health Economics, 13*(8), 793–805. https://doi.org/10.1002/hec.858

50. Jönsson, L., Andreasen, N., Kilander, L., Soininen, H., Waldemar, G., Nygaard, H., Winblad, B., Jönhagen, M. E., Hallikainen, M., & Wimo, A. (2006). Patient- and proxy-reported utility in Alzheimer disease using the EuroQoL. *Alzheimer Disease and Associated Disorders, 20*(1), 49–55. https://doi.org/10.1097/01.wad.0000201851.52707.c9

51. Kaambwa, B., Gill, L., McCaffrey, N., Lancsar, E., Cameron, I. D., Crotty, M., Gray, L., & Ratcliffe, J. (2015). An empirical comparison of the OPQoL-Brief, EQ-5D-3 L and ASCOT in a community dwelling population of older people. *Health and Quality of Life Outcomes, 13*(1), 164. https://doi.org/10.1186/s12955-015-0357-7

52. Karlawish, J. H., Zbrozek, A., Kinosian, B., Gregory, A., Ferguson, A., & Glick, H. A. (2008). Preference-based quality of life in patients with Alzheimer's disease. *Alzheimer's & Dementia, 4*(3), 193–202. https://doi.org/10.1016/j.jalz.2007.11.019

53. Kim, S.-K., Kim, K.-H., Kim, S.-H., Yoo, S.-J., & Jeong, Y.-W. (2019). Health-related quality of life in adult males with lower urinary tract symptoms. *Quality of Life Research, 28*(9), 2419–2428. https://doi.org/10.1007/s11136-019-02205-w

54. Kunz, S. (2010). Psychometric properties of the EQ-5D in a study of people with mild to moderate dementia. *Quality of Life Research, 19*(3), 425–434. https://doi.org/10.1007/s11136-010-9600-1

55. Lutomski, J. E., Krabbe, P. F., Bleijenberg, N., Blom, J., Kempen, G. I., MacNeil-Vroomen, J., Muntinga, M. E., Steyerburg, E., Olde-Rikkert, M. G., & Melis, R. J. (2017). Measurement properties of the EQ-5D across four major geriatric conditions: Findings from TOPICS-MDS. *Health and Quality of Life Outcomes, 15*(1), 45. https://doi.org/10.1186/s12955-017-0616-x

56. Malkin, A. G., Goldstein, J. E., Perlmutter, M. S., & Massof, R. W. (2013). Responsiveness of the EQ-5D to the effects of low vision rehabilitation. *Optometry and vision science, 90*(8), 799–805. https://doi.org/10.1097/opx.0000000000000005

57. Michalowsky, B., Xie, F., Kohlmann, T., Gräske, J., Wübbeler, M., Thyrian, J. R., & Hoffmann, W. (2020). Acceptability and validity of the EQ-5D in patients living with dementia. *Value in Health, 23*(6), 760–767. https://doi.org/10.1016/j.jval.2020.01.022

58. Naglie, G., Tomlinson, G., Tansey, C., Irvine, J., Ritvo, P., Black, S. E., Freedman, M., Silberfeld, M., & Krahn, M. (2006). Utility-based quality of life measures in Alzheimer's disease. *Quality of Life Research, 15*(4), 631–643. https://doi.org/10.1007/s11136-005-4364-8

59. Olerud, P., Tidermark, J., Ponzer, S., Ahrengart, L., & Bergström, G. (2011). Responsiveness of the EQ-5D in patients with proximal humeral fractures. *Journal of Shoulder and Elbow Surgery, 20*(8), 1200–1206. https://doi.org/10.1016/j.jse.2011.06.010

60. Orgeta, V., Edwards, R. T., Hounsome, B., Orrell, M., & Woods, B. (2015). The use of the EQ-5D as a measure of health-related quality of life in people with dementia and their carers. *Quality of Life Research, 24*(2), 315–324. https://doi.org/10.1007/s11136-014-0770-0

61. Parsons, N., Griffin, X. L., Achten, J., & Costa, M. L. (2014). Outcome assessment after hip fracture: Is EQ-5D the answer? *Bone Joint Res, 3*(3), 69–75. https://doi.org/10.1302/2046-3758.33.2000250

62. Pérez-Ros, P., & Martínez-Arnau, F. M. (2020). EQ-5D-3L for assessing quality of life in older nursing home residents with cognitive impairment. *Life, 10*(7), 100.

63. Pérez-Ros, P., Vila-Candel, R., Martin-Utrilla, S., & Martínez-Arnau, F. M. (2020). Health-related quality of life in community-dwelling older people with cognitive impairment: EQ-5D-3L measurement properties. *Journal of Alzheimer's Disease, 77*(4), 1523–1532. https://doi.org/10.3233/jad-200806

64. Sanchez-Arenas, R., Vargas-Alarcon, G., Sanchez-Garcia, S., Garcia-Peña, C., Gutierrez-Gutierrez, L., Grijalva, I., Garcia-Dominguez, A., & Juárez-Cedillo, T. (2014). Value of EQ-5D in Mexican city older population with and without dementia (SADEM study). *International Journal of Geriatric Psychiatry, 29*(5), 478–488. https://doi.org/10.1002/gps.4030

65. Tidermark, J., Bergström, G., Svensson, O., Törnkvist, H., & Ponzer, S. (2003). Responsiveness of the EuroQol (EQ 5-D) and the SF-36 in elderly patients with displaced femoral neck fractures. *Quality of Life Research, 12*(8), 1069–1079. https://doi.org/10.1023/a:1026193812514

66. Tidermark, J., & Bergström, G. (2007). Responsiveness of the EuroQol (EQ-5D) and the Nottingham health profile (NHP) in elderly patients with femoral neck fractures. *Quality of Life Research, 16*(2), 321–330. https://doi.org/10.1007/s11136-006-9004-4

67. van Leeuwen, K. M., Bosmans, J. E., Jansen, A. P., Hoogendijk, E. O., van Tulder, M. W., van der Horst, H. E., & Ostelo, R. W. (2015). Comparing measurement properties of the EQ-5D-3L, ICECAP-O, and ASCOT in frail older adults. *Value in Health, 18*(1), 35–43. https://doi.org/10.1016/j.jval.2014.09.006

68. Walters, S. J., Morrell, C. J., & Dixon, S. (1999). Measuring health-related quality of life in patients with venous leg ulcers. *Quality of Life Research, 8*(4), 327–336. https://doi.org/10.1023/a:1008992006845

69. Aguirre, E., Kang, S., Hoare, Z., Edwards, R. T., & Orrell, M. (2016). How does the EQ-5D perform when measuring quality of life in dementia against two other dementia-specific outcome measures? *Quality of Life Research, 25*(1), 45–49. https://doi.org/10.1007/s11136-015-1065-9

70. Bjerk, M., Brovold, T., Davis, J. C., & Bergland, A. (2019). Evaluating a falls prevention intervention in older home care recipients: A comparison of SF-6D and EQ-5D. *Quality of Life Research, 28*(12), 3187–3195. https://doi.org/10.1007/s11136-019-02258-x

71. Cheng, Q., Kularatna, S., Lee, X. J., Graves, N., & Pacella, R. E. (2019). Comparison of EQ-5D-5L and SPVU-5D for measuring quality of life in patients with venous leg ulcers in an Australian setting. *Quality of Life Research, 28*(7), 1903–1911. https://doi.org/10.1007/s11136-019-02128-6

72. Easton, T., Milte, R., Crotty, M., & Ratcliffe, J. (2018). An empirical comparison of the measurement properties of the EQ-5D-5L, DEMQOL-U and DEMQOL-Proxy-U for older people in residential care. *Quality of Life Research, 27*(5), 1283–1294. https://doi.org/10.1007/s11136-017-1777-0

73. Griffiths, A. W., Smith, S. J., Martin, A., Meads, D., Kelley, R., & Surr, C. A. (2020). Exploring self-report and proxy-report quality-of-life measures for people living with dementia in care homes. *Quality of Life Research, 29*(2), 463–472. https://doi.org/10.1007/s11136-019-02333-3

74. Martin, A., Meads, D., Griffiths, A. W., & Surr, C. A. (2019). How should we capture health state utility in dementia? Comparisons of DEMQOL-proxy-U and of self- and proxy-completed EQ-5D-5L. *Value in Health, 22*(12), 1417–1426. https://doi.org/10.1016/j.jval.2019.07.002

75. Nikolova, S., Hulme, C., West, R., Pendleton, N., Heaven, A., Bower, P., Humphrey, S., Farrin, A., Cundill, B., Hawkins, R., & Clegg, A. (2020). Normative estimates and agreement between 2 measures of health-related quality of life in older people with frailty: Findings from the community ageing research 75+ cohort. *Value in Health, 23*(8), 1056–1062. https://doi.org/10.1016/j.jval.2020.04.1830

76. Ratcliffe, J., Flint, T., Easton, T., Killington, M., Cameron, I., Davies, O., Whitehead, C., Kurrle, S., Miller, M., Liu, E., & Crotty, M. (2017). An Empirical comparison of the EQ-5D-5L, DEMQOL-U and DEMQOL-proxy-U in a post-hospitalisation population of frail older people living in residential aged care. *Applied Health Economics and Health Policy, 15*(3), 399–412. https://doi.org/10.1007/s40258-016-0293-7

77. Janssen, M. F., Lubetkin, E. I., Sekhobo, J. P., & Pickard, A. S. (2011). The use of the EQ-5D preference-based health status measure in adults with type 2 diabetes mellitus. *Diabetic Medicine, 28*(4), 395–413. https://doi.org/10.1111/j.1464-5491.2010.03136.x

78. Cameron, L. J., Wales, K., Casey, A., Pike, S., Jolliffe, L., Schneider, E. J., Christie, L. J., Ratcliffe, J., & Lannin, N. A. (2021). Self-reported quality of life following stroke: A systematic review of instruments with a focus on their psychometric properties. *Quality of Life Research*. https://doi.org/10.1007/s11136-021-02944-9

79. Feng, Y.-S., Kohlmann, T., Janssen, M. F., & Buchholz, I. (2021). Psychometric properties of the EQ-5D-5L: A systematic review of the literature. *Quality of Life Research, 30*(3), 647–673. https://doi.org/10.1007/s11136-020-02688-y

80. Devlin, N., Parkin, D., & Janssen, B. (2020). Advanced topics. *Methods for analysing and reporting EQ-5D data* (pp. 87–98). Springer International Publishing.

81. Davis, J. C., Khan, K. M., Hsu, C. L., Chan, P., Cook, W. L., Dian, L., & Liu-Ambrose, T. (2020). Action seniors! cost-effectiveness analysis of a secondary falls prevention strategy among community-dwelling older fallers. *Journal of the American Geriatrics Society, 68*(9), 1988–1997. https://doi.org/10.1111/jgs.16476

82. Hewitt, J., Saing, S., Goodall, S., Henwood, T., Clemson, L., & Refshauge, K. (2019). An economic evaluation of the SUNBEAM programme: A falls-prevention randomized controlled trial in residential aged care. *Clinical Rehabilitation, 33*(3), 524–534. https://doi.org/10.1177/0269215518808051

83. Robertson, M. C., Campbell, A. J., Gardner, M. M., & Devlin, N. (2002). Preventing injuries in older people by preventing falls: A meta-analysis of individual-level data. *Journal of the American Geriatrics Society, 50*(5), 905–911. https://doi.org/10.1046/j.1532-5415.2002.50218.x

84. Buchholz, I., Janssen, M. F., Kohlmann, T., & Feng, Y.-S. (2018). A systematic review of studies comparing the measurement properties of the three-level and five-level versions of the EQ-5D. *PharmacoEconomics, 36*(6), 645–661. https://doi.org/10.1007/s40273-018-0642-5

85. Janssen, M. F., Bonsel, G. J., & Luo, N. (2018). Is EQ-5D-5L better than EQ-5D-3L? A head-to-head comparison of descriptive systems and value sets from seven countries. *PharmacoEconomics, 36*(6), 675–697. https://doi.org/10.1007/s40273-018-0623-8

86. Grewal, I., Lewis, J., Flynn, T., Brown, J., Bond, J., & Coast, J. (2006). Developing attributes for a generic quality of life measure for older people: Preferences or capabilities? *Social Science and Medicine, 62*(8), 1891–1901. https://doi.org/10.1016/j.socscimed.2005.08.023

87. Power, M., Quinn, K., & Schmidt, S. (2005). Development of the WHOQOL-old module. *Quality of Life Research, 14*(10), 2197–2214. https://doi.org/10.1007/s11136-005-7380-9

88. Cella, D., Choi, S. W., Condon, D. M., Schalet, B., Hays, R. D., Rothrock, N. E., Yount, S., Cook, K. F., Gershon, R. C., Amtmann, D., DeWalt, D. A., Pilkonis, P. A., Stone, A. A., Weinfurt, K., & Reeve, B. B. (2019). PROMIS® adult health profiles: Efficient short-form measures of seven health domains. *Value in Health, 22*(5), 537–544. https://doi.org/10.1016/j.jval.2019.02.004

89. Dewitt, B., Feeny, D., Fischhoff, B., Cella, D., Hays, R. D., Hess, R., Pilkonis, P. A., Revicki, D. A., Roberts, M. S., Tsevat, J., Yu, L., & Hanmer, J. (2018). Estimation of a preference-based summary score for the patient-reported outcomes measurement information system: The PROMIS(®)-preference (PROPr) scoring system. *Medical Decision Making, 38*(6), 683–698. https://doi.org/10.1177/0272989x18776637

90. Hanmer, J., Cella, D., Feeny, D., Fischhoff, B., Hays, R. D., Hess, R., Pilkonis, P. A., Revicki, D., Roberts, M., Tsevat, J., & Yu, L. (2017). Selection of key health domains from PROMIS® for a generic preference-based scoring system. *Quality of Life Research, 26*(12), 3377–3385. https://doi.org/10.1007/s11136-017-1686-2

91. Himmler, S., van Exel, J., & Brouwer, W. (2020). Happy with your capabilities? Valuing ICECAP-O and ICECAP-A states based on experienced utility using subjective well-being data. *Medical Decision Making, 40*(4), 498–510. https://doi.org/10.1177/0272989x20923015

92. Coast, J., Flynn, T. N., Natarajan, L., Sproston, K., Lewis, J., Louviere, J. J., & Peters, T. J. (2008). Valuing the ICECAP capability index for older people. *Social Science and Medicine, 67*(5), 874–882. https://doi.org/10.1016/j.socscimed.2008.05.015

93. Cleland, J., Hutchinson, C., McBain, C., Walker, R., Milte, R., Khadka, J., & Ratcliffe, J. (2021). Developing dimensions for a new preference-based quality of life instrument for older people receiving aged care services in the community. *Quality of Life Research, 30*(2), 555–565. https://doi.org/10.1007/s11136-020-02649-5

94. Ratcliffe, J., Cameron, I., Lancsar, E., Walker, R., Milte, R., Hutchinson, C. L., Swaffer, K., & Parker, S. (2019). Developing a new quality of life instrument with older people for economic evaluation in aged care: Study protocol. *British Medical Journal Open, 9*(5), e028647. https://doi.org/10.1136/bmjopen-2018-028647

95. Geraerds, A. J. L. M., Bonsel, G. J., Janssen, M. F., Finch, A. P., Polinder, S., & Haagsma, J. A. (2021). Methods used to identify, test, and assess impact on preferences of bolt-ons: A systematic review. *Value in Health, 24*(6), 901–916. https://doi.org/10.1016/j.jval.2020.12.011

96. Finch, A. P., Brazier, J. E., Mukuria, C., & Bjorner, J. B. (2017). An exploratory study on using principal-component analysis and confirmatory factor analysis to identify bolt-on dimensions: The EQ-5D case study. *Value in Health, 20*(10), 1362–1375. https://doi.org/10.1016/j.jval.2017.06.002

97. Yang, Y., Rowen, D., Brazier, J., Tsuchiya, A., Young, T., & Longworth, L. (2015). An exploratory study to test the impact on three "bolt-on" items to the EQ-5D. *Value in Health, 18*(1), 52–60. https://doi.org/10.1016/j.jval.2014.09.004

98. Finch, A. P., Brazier, J., & Mukuria, C. (2021). Selecting bolt-on dimensions for the EQ-5D: Testing the impact of hearing, sleep, cognition, energy, and relationships on preferences using pairwise choices. *Medical Decision Making, 41*(1), 89–99. https://doi.org/10.1177/0272989x20969686

99. Finch, A. P., Brazier, J. E., & Mukuria, C. (2019). Selecting bolt-on dimensions for the EQ-5D: Examining their contribution to health-related quality of life. *Value in Health, 22*(1), 50–61. https://doi.org/10.1016/j.jval.2018.07.001

100. Chen, G., & Olsen, J. A. (2020). Filling the psycho-social gap in the EQ-5D: The empirical support for four bolt-on dimensions. *Quality of Life Research, 29*(11), 3119–3129. https://doi.org/10.1007/s11136-020-02576-5

101. Beerens, H. C., Zwakhalen, S. M., Verbeek, H., Ruwaard, D., & Hamers, J. P. (2013). Factors associated with quality of life of people with dementia in long-term care facilities: A systematic review. *International Journal of Nursing Studies, 50*(9), 1259–1270. https://doi.org/10.1016/j.ijnurstu.2013.02.005

102. Jing, W., Willis, R., & Feng, Z. (2016). Factors influencing quality of life of elderly people with dementia and care implications: A systematic review. *Archives of Gerontology and Geriatrics, 66*, 23–41. https://doi.org/10.1016/j.archger.2016.04.009

103. Asakawa, K., Senthilselvan, A., Feeny, D., Johnson, J., & Rolfson, D. (2012). Trajectories of health-related quality of life differ by age among adults: Results from an eight-year longitudinal study. *Journal of Health Economics, 31*(1), 207–218. https://doi.org/10.1016/j.jhealeco.2011.10.002