Journal of Clinical Epidemiology 149 (2022) 1−11

Journal of Clinical Epidemiology

REVIEW ARTICLE

A systematic review highlighting poor quality of evidence for content validity of quality of life instruments in female chronic pelvic pain

Vishalli Ghai[a,b,*], Venkatesh Subramanian[a], Haider Jan[a], Stergios K. Doumouchtsis[a,b,c,d,e], On behalf of CHORUS: An International Collaboration Harmonising Outcomes, Research, and Standards in Urogynaecology and Womens Health

[a]*Department of Obstetrics and Gynaecology, Epsom & St Helier University Hospitals NHS Trust, London KT18 7EG, UK*
[b]*St George's University of London, Crammer Terrace, London SW17 0RE, UK*
[c]*Laboratory of Experimental Surgery and Surgical Research N.S. Christeas, Medical School, National and Kapodistrian University of Athens, Athens, Greece*
[d]*American University of the Caribbean, Pembroke Pines, FL, USA*
[e]*Ross University, School of Medicine, Miramar, FL, USA*

Accepted 12 April 2022; Published online 19 April 2022

Abstract

Objectives: To evaluate the content validity of 19 patient-reported outcome measures (PROMs) used to measure quality of life (QoL) in women with chronic pelvic pain (CPP).

Study Design and Setting: We searched Embase, MEDLINE, PsycINFO databases and Google Scholar from inception to August 2020. We included records describing the development or studies assessing content validity of PROMs. Two reviewers independently assessed the methodological quality of PROMs using the Consensus-based Standards for the Selection of Health Measurement Instruments checklist. Evidence was synthesized for relevance, comprehensiveness, and comprehensibility. Quality of evidence was rated using a modified Grading of Recommendations, Assessment, Development, and Evaluations approach.

Results: PROM development was inadequate for all instruments included in this review. No high-quality evidence ratings were found for relevance, comprehensiveness, and comprehensibility. QoL was measured using generic instruments (68.42%, 13/19) rather than those specific to chronic pain (21.04%, 4/19) or pelvic pain (10.53%, 2/19). Quality of concept elicitation was inadequate for 90% of PROMs. Half of PROMs did not include patients in their development and only 40% were devised using a sample representative of the target population for which the PROM was developed. Cognitive interviews were conducted in one-fifth of PROMs and were mostly of inadequate/doubtful quality.

Conclusion: There is poor quality of evidence for content validity of PROMs used to measure QoL in women with CPP.

Keywords: Chronic pelvic pain; Pain; Women; Outcomes; Content validity; Patient-reported outcome measures; Systematic review; Core outcome sets; COSMINCHORUS

1. Introduction

Chronic pelvic pain (CPP) is a debilitating condition associated with significant long-term morbidity and socioeconomic burden [1,2]. The complexities of pain perception and sensation mean that it is seldom curative. Therefore, clinical efforts are focused on reducing pain intensity and improving health related quality of life (QoL). QOL is defined as physical, psychological, and social domains of health, seen as distinct areas that are influenced by a person's experiences, beliefs, expectations, and perceptions [3].

Funding: None received.

Details of Ethical Approval: This systematic review is based on outcomes published in previous trials. No approval was required from an institutional review board.

Declaration of interest: All authors have no conflicts of interest to declare.

* Corresponding author. Department of Obstetrics and Gynaecology, Epsom & St Helier's University Hospitals NHS Trust, Dorking Road, London, UK KT18 7EG. Tel.: +44-1372-735-735.

E-mail address: Vishalli.ghai25@googlemail.com (V. Ghai).

https://doi.org/10.1016/j.jclinepi.2022.04.016

What is new?

Key findings

- We identified, summarized, and graded quality of evidence supporting content validity of 19 patient-reported outcome measures (PROMs) reporting quality of life (QoL) outcomes in women with CPP.
- This systematic review has shown poor quality evidence for content validity of PROMs measuring QoL in women with CPP including inadequate PROM development in areas such as concept elicitation and the use of cognitive interviews.

What this add to what is known?

- This is the first systematic review to implement the Consensus-based Standards for the Selection of Health Measurement Instruments (COSMIN) criteria to assess the content validity of PROMs reporting QoL outcomes in women with CPP.

What is the implication and what should change now?

- Findings of this review are concerning for clinicians. It is essential that high content validity instruments are used to generate relevant and meaningful data as it may influence decisions made by health professionals and patients.
- Our evaluation of content validity of PROMs assessing QoL in women with CPP may be used subsequently to recommend instruments to measure core outcomes in core outcome sets in female CPP.

The measurement of QoL is considered, an important outcome domain among researchers and clinicians in clinical trials and a priority among women with CPP [4,5]. Several measurement instruments are available to measure QoL. The selection of adequate instruments is determined by validity, which is the extent to which an instrument accurately measures what it is supposed to measure. The Consensus-based Standards for the Selection of Health Measurement Instruments (COSMIN) taxonomy divides validity into five subdomains [6]. Content validity is the first measurement property considered when choosing a patient-reported outcome measure (PROM) and is described as the degree to which the content of an instrument is an adequate reflection of the construct to be measured [7]. It refers to the relevance, comprehensiveness, and comprehensibility of a PROM with respect to the construct, target population, and context of use. Content validity influences other measurement properties. For example, poor content validity can impact the responsiveness of a PROM.

2. Objectives

There are no systematic reviews available on content validity of PROMs measuring QoL in women with CPP using a standardized COSMIN methodology. Previous reviews published on PROMs in benign gynaecology including CPP and endometriosis have been limited [8−11]. Results have been descriptive, presented basic psychometric properties and not performed using a standardized COSMIN methodology.

A thorough assessment of content validity should not only include studies evaluating content validity in the population of interest but also original development studies and the content of the instrument itself. The COSMIN initiative has developed methodological guidance describing what constitutes sufficient content validity including a method to integrate methodological quality and results into an evidence synthesis rating system [12].

This systematic review has applied a COSMIN methodology to evaluate PROMs used to measure QoL in women with CPP. As an important element of this evaluation process, we considered content validity to be a parameter of particular relevance for the reasons described above.

This review was performed by a working group of CHORUS, an International Collaboration for Harmonising Outcomes in Research, and Standards in Urogynaecology and Women's Health (https://i-chorus.org) and is part of a wider initiative to develop core outcome sets in CPP. In depth, assessment of outcome measures for suitability prior to consideration for inclusion in a core outcome measure set has been recommended and this study aims to contribute to this process.

3. Materials and methods

This systematic review was registered with the Core Outcomes Measures in Effectiveness Trials (COMET) initiative register, number 981 and with the International Prospective Register of Systematic Reviews (PROSPERO), registration number CRD42019134858. We have performed a secondary analysis of our previous findings on the variation of outcomes and applied outcome measures in CPP trials [4]. Consequently, we have employed additional methodology for the purpose of this specific review which was not stated in the initial protocol registered on PRSOPERO. These include those related to our search strategy, evidence synthesis, and risk of bias. For example, our search strategy included additional databases such as PsycInfo and COSMIN and data sources such as Google Scholar. We performed an evidence synthesis and a risk of bias assessment in accordance with the standardized approach as recommended by the COSMIN guidelines in relation to systematic reviews evaluating PROMs [12−14].

3.1. Study design

The design of the present systematic review was based on the Preferred Reporting Items for Systematic

Reviews and Meta-Analyses (PRISMA) guideline [15] (Appendix A).

Our inventory of PROMs measuring QoL in women with CPP was informed by a previous systematic review reporting all outcomes and outcome measures in effectiveness trials assessing interventions in idiopathic CPP [4]. A comprehensive literature search was undertaken using Cochrane Central Register of Controlled Trials (CENTRAL), Embase, and MEDLINE databases. Searches were performed from database inception to September 2019 using the following Medical Subject Heading search terms: ''CPP,'' ''pelvic pain,'' and ''idiopathic CPP''. We only included randomized control trials assessing the effectiveness of interventions for CPP. The population of interest included women aged more than 18 years with CPP. We included all studies investigating psychological therapies, medical and surgical interventions with existing treatments, or placebo regimes. We excluded studies in languages other than English, pilot studies, nonrandomized studies, retrospective studies, case series, and case reports. We identified 48 measurement instruments, of which 15 PROMs assessed QoL including becks depression inventory, brief pain inventory−interference subscale, Endometriosis Health Profile 30 (EHP 30), EuroQoL 5D (EQ-5D), fear-avoidance beliefs questionnaire (FABQ), General Health Questionnaire (GHQ), Hospital Anxiety and Depression Scale (HADS), inventory of interpersonal problems (IIP), multidimensional pain inventory (MPI), Oswestry Disability Index (ODI), Pain Beliefs and Perception Inventory (PBPI), Sexual Activity Questionnaire (SAQ), short form health survey (SF 36), short from health survey 12 (SF 12), and the World Health Organization Quality of Life assessment (WHOQoL).

For the purpose of this review, we included all versions of a PROM which resulted in a total of 19 PROMs: becks depression inventory 2.0, brief pain inventory, EHP 30, EHP 5, EQ-5D 3L, EQ-5D 5L, FABQ, HADS, IIP-64, IIP-32, MPI, ODI 1.0, ODI 2.1a, PBPI, SAQ, SF 36, SF 12, WHOQoL 100, and WHOQoL Bref. The GHQ was not included as we could obtain the user manual or questionnaire from the publisher.

3.2. Search strategy

A comprehensive literature search was conducted using Embase, MEDLINE, and PsycINFO databases from inception to August 2020. The search strategy consisted of three groups of search terms combined with the Boolean operator ''AND'' (1): instrument names (2), CPP, and (3) measurement properties to identify and evaluate evidence for this current systematic review. A previously developed search filter used to retrieve studies on measurement properties in PubMed was adapted for all other databases [16] (Appendix B for search strategy). Google scholar was also searched (search date November 17, 2020) using the name of PROMs, the first 100 webpages for each PROM were screened for inclusion. Citation tracking of eligible records was also conducted. Results of searches were combined and duplicates removed in Endnote 20.

3.3. Selection of studies

Any report (i.e., book, online article) presenting the development of the 19 PROMs was included for the assessment of content validity [12]. Content validity studies were eligible for inclusion if they were full-text original articles, about women with nonspecific CPP or professionals to assess the relevance, comprehensiveness, and comprehensibility of the content of at least one of the PROMs [12]. Studies that included women with mixed pathologies were included if at least 75% of the total sample had nonspecific CPP. Studies on cross-cultural adaptation were included as content validity studies if they performed a pilot study of the adapted questionnaire, in which its comprehensibility was assessed in patients with nonspecific CPP [17].

Two independent researchers (V.G. and V.S.) screened for potentially eligible studies by examining initially titles and subsequently abstracts of the identified studies. Full-text articles were retrieved for abstracts meeting the inclusion criteria or in cases when information in the abstract was incomplete or unclear. Full-text articles were reviewed and discrepancies regarding suitability for inclusion were resolved by discussion with the senior author (S.K.D.). Results of the study selection process were summarized in a flow chart including reasons for excluding the full text (Fig. S1 for the flow chart and Appendix D for the list of excluded records). References were managed by Endnote 20.

3.4. Data extraction

A standardized data extraction form was used [14]. The following information was extracted: characteristics of the PROM (i.e., construct, target population, intended context of use, mode/time of administration, number of scales/items/response options, recall period, range of scores, available translations, and access fee) and characteristics of the development study (conceptual framework, language, and patient involvement). Data extraction was performed by two researchers independently (V.G. and V.S.) and in the case of disagreement a consensus was reached by discussion with a third senior reviewer (S.K.D.).

3.5. Quality assessment

The methodological quality of PROM development and content validity was assessed using the COSMIN risk of bias checklist.

PROM development was evaluated using 35 standards divided across two areas (1): quality of PROM design including the concept elicitation study for item generation and (2) quality of the cognitive interview study to assess

Table 1. Characteristics of included PROMS

PROM[a] (reference to first article)	Construct(s)	Target population PROM developed for
BDI II [19,20]	Indicator of the presence and degree of depressive symptoms	Patients diagnosed with depression
BPI- Pain Interference subscale [21]	Measure of the severity and impact of cancer-related pain on functioning	Patients with cancer related pain
EHP 30 [22]	Assessment of health related quality of life, ''encompassing physical, psychologic, and social aspects'', of women with endometriosis	Women with endometriosis
EHP 5 [22,23]	Assumed same as EHP-30	Assumed same as EHP-30
EQ-5D-3L [24,25]	Generic measure of health related quality of life, no definition given.	Nondisease specific.
EQ-5D-5L [26,27] (2009)	Assumed same as EQ-5D-3L	Assumed same as EQ-5D-3L
FABQ [28]	Measure of patients' fear of pain and consequent avoidance of physical activity and long-term disability.	Patients with chronic lower back pain
HADS [29] (1983)	Measure to detect depression and anxiety	Patients in hospital clinics
IIP 64 [30,31]	Measure of distress and determining source of interpersonal difficulties, by assessing eight domains: Domineering/controlling, vindictive/self-centred, cold/distant, socially inhibited, nonassertive, overly accommodating, self-sacrificing, and intrusive/needy.[b]	Patients attending psychotherapy reporting interpersonal difficulties
IIP 32 [31,32]	Assumed same as IIP 64	Assumed same as IIP 64
MPI [33]	Measure of the subjective distress caused by pain and impact of pain on patients' lives	Patient with chronic pain (men and women)
ODI 1.0 [34]	Disability defined as the limitations of a patient's performance compared with that of a fit person	Patients with acute or chronic lower back pain
ODI 2.1a [35]	Assumed same as Oswestry Disability Index 1.0	Assumed same as Oswestry Disability Index 1.0
PBPI [36,37]	Measure of pain beliefs, assessing four domains: mystery, self-blame, constancy, and permanence[b]	Injured workers (men and women) receiving compensation with chronic pain as a result of injury at work, not defined.
SAQ [38]	Measure of sexual function, no definition given	Women on long-term Tamoxifen with a high risk of developing breast cancer.
SF-36 [39,40]	Generic health, eight concepts, assessing physical functioning, social and role functioning, mental health, general health, perceptions, bodily pain, and vitality[b]	General and patient population
SF-12 [41]	Assumed same as SF-36	Assumed same as SF-36
WHOQoL-100 [42,43]	Generic measure of quality of life cross-culturally (definition given), six domains identified as core aspect of quality of life cross-culturally: physical, psychological, level of independence, social relationships, environment, and personal beliefs/spiritualty[b]	Patient groups in both developing and developed countries
WHOQoL-Bref [42,44]	Generic measure of quality of life cross-culturally (definition given), four domains identified as core aspect of quality of life cross-culturally: physical, psychological, social relationships, and environment[b]	Assumed same as WHOQol-100

Abbreviations: BDI, becks depression inventory; BPI, brief pain inventory; EHP 30, endometriosis health profile 30; EHP-5, endometriosis health profile 5; EQ-5D 3L, EuroQoL 5D 3L; EQ-5D 5L, EuroQoL 5D 5L; FABQ, fears avoidance beliefs questionnaire; HADS, hospital anxiety and depression scale; IIP 64, inventory of interpersonal problems 64; IIP 32, inventory of interpersonal problems 32; MPI, multidimensional pain inventory; ODI 1.0, oswestry disability index 1.0; ODI 2.1a, oswestry disability index 2.1a; PBPQ, pain beliefs and perception questionnaire; PROM, patient-reported outcome measures; QoL, quality of life; SAQ, sexual activity questionnaire; SF 36, short form survey 36; SF 12, short form survey 12; WHOQoL, World Health Organization Quality of Life Questionnaire; PROM, patient reported outcome measures.

[a] Each version of a PROM is considered a separate PROM.

[b] These domains/concept have been defined, please refer to the reference.

Context PROM developed for	Mode of administration (e.g., self-report, interview-based, proxy report, etc.)	Language (s) (country) of development	Available translations	Access fee	Patient Involvement in concept elicitation
Clinical practice and research	Self-administered	English (US)	Yes	Yes	Yes
Clinical practice, clinical trials, epidemiological research	Interview-based	English (US)	Yes	Yes	Yes
Clinical practice and research	Self-administered	English (UK)	Yes	Yes	Yes
Assumed same as EHP-30	Assumed same as EHP-30	English (UK)	Yes	Yes	Yes
Large-scale surveys of community	Self-administered, interview-based	Dutch English (UK) Finnish Norwegian Swedish	Yes	Yes	None
Assumed same as EQ-5D-3L	Assumed same as EQ-5D-3L	English (UK) Spanish (Spain)	Yes	Yes	Yes
Clinical practice and research	Self-administered	English (UK)	Yes	Not stated	Yes
Clinical practice and research	Self-administered	English (UK)	Yes	Yes	None
Clinical practice, research	Self-administered, interview-based	English (US)	Yes	Yes	Yes
Assumed same as IIP 64	Self-administered, interview-based	English (US)	Yes	Yes	Yes
Clinical practice and research	Self-administered, interview- based	English (US)	Yes	None	None
Clinical response to treatment	Self-administered	English (UK)	Yes	Yes	None
Assumed same as Oswestry Disability Index 1.0	Assumed same as Oswestry Disability Index 1.0	English (UK)	Yes	Yes	None
Clinical practice, research	Self-administered	English (US)	No	Not stated	**Yes**
Unclear context, implied for clinical trials	Self-administered	English (UK)	No	Not stated	None
Clinical practice, research, health policy evaluations and general population health survey	Self-administered, interview administration	English (US)	Yes	None	None
Assumed same as SF-36	Assumed same as SF-36	English (US)	Yes	None	None
Clinical practice, clinical trials, epidemiological research, health policy, and service evaluation	Self-administered, interview-based	Various languages -more than 30	Yes	None	Yes
Assumed same as WHOQol-100	Assumed same as WHOQol-100	Various languages -more than 30	Yes	Yes	**Yes**

relevance, comprehensiveness, and comprehensibility of PROM items. Each standard was rated using a four-point scale: ''very good,'' ''adequate,'' ''doubtful,'' or ''inadequate''.

A second set of COSMIN standards was used to assess the methodological quality of content validity studies. A total of 31 standards evaluated studies which reported responsiveness, comprehensiveness, or comprehensibility by patients or professionals. Each standard was rated using a four-point scale: ''very good,'' ''adequate,'' ''doubtful,'' or ''inadequate''.

Total scores were calculated for both parts of the PROM development study (quality of PROM design and quality of cognitive interview study) and for each aspect of the methodological quality of the content validity studies (relevance, comprehensiveness, and comprehensibility). Total scores given to each box/part of box were determined using the lowest grade of any standard in that box/part of box (i.e., ''the worst score counts'' principle).

We also searched the COSMIN database to identify previous studies which evaluated the quality of PROM development of instruments included in this review. We used these findings of PROM development to support our evaluation.

A quality assessment was performed by two researchers independently (V.G. and V.S.) and in the case of disagreement a consensus was reached by discussion with a third senior reviewer (S.K.D.).

3.6. Evidence synthesis

First, the PROM development study, content validity studies, and the content of the PROM were rated against the 10 criteria of good content validity [12]. There are five criteria for relevance, one criterion for comprehensiveness, and four criteria for comprehensibility. Each criterion was scored sufficient (+), insufficient (−), or indeterminate (?).

Second, results of the PROM development study, content validity studies, and reviewer ratings of PROM content were qualitatively summarized and compared against the criteria of good content validity. Overall ratings for relevance, comprehensiveness, comprehensibility, and overall content validity were determined for each PROM. Ratings were sufficient (+), insufficient (−), or inconsistent (±).

Finally, a modified Grading of Recommendations, Assessment, Development, and Evaluations approach was applied to assess the quality of evidence [18]. Factors considered include study quality, consistency of results across studies, and indirectness. The quality of evidence was graded as high, moderate, low, and very low.

Data synthesis was performed by two researchers independently (V.G. and V.S.) and in the case of disagreement a consensus was reached by discussion with a third senior reviewer (S.K.D.).

3.7. Data analysis

We used a descriptive method of analysis to produce overall ratings for relevance, comprehensiveness, comprehensibility, and content validity per PROM including a quality assessment of evidence.

3.8. Patient and public involvement

There has been no patient involvement as this study is a systematic review of existing research.

4. Results

The literature search was conducted on August 30, 2020 and we identified 475 titles and abstracts. We screened 307 titles and abstracts following the exclusion of 168 duplicate records. Five records were identified from our literature search (Fig. S1). Further 27 records were identified from our search on the COSMIN database, Google Scholar, and citation tracking. In total, we included 32 records focussing on PROM development of 19 PROMs and one study assessing content validity involving patients of a single PROM (Appendix C for a list of included records and Appendix D for a list of excluded full-text records).

4.1. Quality of patient-reported outcome measure development study

Table 1 presents a summary of the development studies describing the construct definition, target population, context of use, and patient involvement of 19 PROMs (an online table is available (Table S1) and presents further details of PROM characteristics).

Four PROMs (BPI, FABQ, MPI, and PBPI) were developed to assess QoL in patients with chronic pain. Only two PROMs (EHP 30 and EHP 5) were designed specifically to assess QoL in women with CPP; however, this was secondary to endometriosis. The remaining 13 PROMs were developed generically to assess the QoL among patients with various health conditions.

Overall, PROM development was considered inadequate for all instruments included in this review. Almost half of PROMs (8/19, 42.11%) did not involve patients in their development (EQ5D-3L, HADS, MPI, ODI 1.0, ODI 2.1a, SAQ, SF 12, and SF 36) (Table 1). Concept elicitation was deemed inadequate for 17 PROMs (Table 2). The other two PROMs, BPI and EQ5D-5L, were considered doubtful as it was unclear whether the patients included were representative of the target population. The number and characteristics of patients in the development were not reported. Eight PROMs (8/19, 42.11%) failed to conduct development studies in a sample representing the target population for which the PROM was developed (EQ-5D 3L, HADS, MPI, ODI 1.0, ODI 2.1a, SAQ, SF 12, and SF 36) (Table 2).

Only four PROMs featured cognitive interviews with patients in their development process (EHP-30, EQ5D-5L, WHOQol, and WHOQol-Bref) (Table 2). Cognitive interviews were inadequate for the EQ 5D-5L because the final form of the questionnaire was not tested. Cognitive interviews were of doubtful quality for the following PROM's: EHP-30, WHOQol, and WHOQol-Bref. For the EHP-30 no details were provided regarding the methods used to assess the comprehensibility and comprehensiveness. Cognitive interviews for the WHOQoL and WHOQoL-Bref did examine comprehensibility; however, limited details regarding assessment were provided. It was unclear whether comprehensiveness was evaluated for the WHOQoL and WHOQoL-Bref.

4.2. Content validity studies

We only identified one study evaluating content validity of a single PROM−SF 36 [49]. This study involved 105 female patients presenting with CPP and assessed relevance, comprehensibility, and comprehensiveness. This study was of doubtful quality as it was unclear which aspect was assessed. In addition, the study did not report the use of an interview/topic guide, trained moderators/interviewers, and whether two independent researchers conducted the analysis. No content validity studies were found including professionals.

4.3. Evidence synthesis

No high-quality evidence was available for the 19 PROMs included in this review (Table 3) . All PROMs had low or very low quality of evidence ratings for relevance, comprehensiveness, and comprehensibility. Four PROMs achieved a low quality of evidence rating for relevance (SF 36, EHP 30, EHP 5, and MPI). Quality of evidence on comprehensiveness was rated as low for two instruments (SF 36 and EHP 30). Only the SF 36 attained a low quality of evidence rating for comprehensibility. Based on these results, it is not possible to establish which PROM has the best content validity for QoL in women with CPP.

5. Discussion

5.1. Main findings

Overall, PROM development was inadequate for all instruments used to assess QoL in women with CPP. No high-quality evidence ratings were found for relevance, comprehensiveness, and comprehensibility. QoL was measured using generic instruments (68.42%, 13/19) rather than those specific to chronic pain (21.04%, 4/19) or pelvic pain (10.53%, 2/19). Quality of concept elicitation was inadequate for 90% of PROMs with most failing to involve patients in their development or a sample representative of the target population for which the PROM was developed.

Only a fifth of PROMs were developed using cognitive interviews assessing comprehensiveness and comprehensibility. We identified one content validity study assessing the relevance, comprehensiveness, and comprehensibility of the SF 36 within a CPP population.

5.2. Strengths and limitations

This is the first systematic review to implement COSMIN criteria to assess the content validity of PROMs used to evaluate QoL in women with CPP. In our evaluation, we considered the methodological quality, the development process, findings of the content validity study, and content of the instrument itself. We used robust and reproducible methods, which have been successfully implemented in studies evaluating content validity of PROMs in various medical specialities [45,50,51].

This study has limitations. The validity of an instrument is dependent on the interpretation of instrument scores in a given application [45]. Therefore, our findings may not be generalizable to every context. In addition, the current perspective on validity differs and focuses only on inferences, claims, or decisions made, based on instrument scores rather than the instruments itself [45]. We limited the inclusion of PROMs to those from effectiveness trials only. There may be a possibility that this is not an exhaustive list of all PROMs reporting QoL in CPP. However, this systematic review is part of a wider initiative to produce a COS which will stipulate a minimum set of outcomes for effectiveness trials to report. Therefore, our study findings will support and contribute to the evaluation, selection, and recommendation of appropriate instruments to measure these core outcomes. Although we aimed to include studies describing cross-cultural adaptations of PROMs, we found no studies specific to women with idiopathic CPP. Consequently, our results were limited to studies written in the English language. Content validity is one of many parameters to assess the quality of measurement instruments. The focus of our study may appear limited; however, based on our results, further analyses may be conducted to assess additional parameters.

5.3. Interpretation

Findings of this review are concerning for clinicians who routinely use PROMs in their clinical practice. We demonstrated that frequently used PROMs reporting QoL outcomes lack content validity. It is essential that high content validity instruments are used to generate data which are relevant and meaningful as they may influence decisions made by health professionals and patients regarding current or future treatment options. For example, the European Medicines Agency has acknowledged the role of patients' perspective including the impact of anticancer medication affecting their wellbeing and daily life. The collection of PROMs in this instance is an important aspect of evaluating

Table 2. Content validity assessment for QoL instruments for female CPP

PROM	Reference	Concept elicitation study; quality[a]	Concept elicitation study; patient involvement	Cognitive study performed	Cognitive study; quality[a]	Overall quality of PROM development study[a]
BDI 2.0	[19,20]	Inadequate	Yes	No	n/a	Inadequate
BPI	[45,21]	Doubtful	Yes	No	n/a	Inadequate
EHP 30	[22]	Inadequate	Yes	Yes	Doubtful	Inadequate
EHP 5	[22,23]	Inadequate	Yes	No	n/a	Inadequate
EQ-5D 3L	[3]	Inadequate	None	No	n/a	Inadequate
EQ-5D 5L	[27,46]	Doubtful	Yes	Yes	Inadequate	Inadequate
FABQ	[28]	Inadequate	Yes	No	n/a	Inadequate
HADS	[29,46]	Inadequate	None	No	n/a	Inadequate
IIP-64	[30,31,47,48]	Inadequate	Yes	No	n/a	Inadequate
IIP-32	[30–32,48]	Inadequate	Yes	No	n/a	Inadequate
MPI	[45,33]	Inadequate	None	No	n/a	Inadequate
ODI 1.0	[45]	Inadequate	None	No	n/a	Inadequate
ODI 2.1a	[45]	Inadequate	None	No	n/a	Inadequate
PBPQ	[36]	Inadequate	Yes	No	n/a	Inadequate
SAQ	[38]	Inadequate	None	No	n/a	Inadequate
SF 36	[45,39]	Inadequate	None	No	n/a	Inadequate
SF 12	[45,41]	Inadequate	None	No	n/a	Inadequate
WHOQoL	[42,43]	Inadequate	Yes	Yes	Doubtful	Inadequate
WHOQoL-Bref	[42,44]	Inadequate	Yes	Yes	Doubtful	Inadequate

Abbreviations: BDI, becks depression inventory; BPI, brief pain inventory; CPP, chronic pelvic pain; EHP 30, endometriosis health profile 30; EHP-5, endometriosis health profile 5; EQ-5D 3L, EuroQoL 5D 3L; EQ-5D 5L, EuroQoL 5D 5L; FABQ, fears avoidance beliefs questionnaire; HADS, hospital anxiety and depression scale; IIP 64, inventory of interpersonal problems 64; IIP 32, inventory of interpersonal problems 32; MPI, multidimensional pain inventory; ODI 1.0, oswestry disability index 1.0; ODI 2.1a, oswestry disability index 2.1a; PBPQ, pain beliefs and perception questionnaire; QoL, quality of life; SAQ, sexual activity questionnaire; SF 36, short form survey 36; SF 12, short form survey 12; WHO-QoL, World Health Organization Quality of Life Questionnaire.

[a] Quality rated as very good, adequate, doubtful, inadequate, or not applicable.

clinical benefits and efficacy of new drugs which are not gained from objective or clinical assessments [52].

High content validity of PROMs is ensured by items generated during concept elicitation reflecting what is important to patients. This can be accomplished by undertaking interviews/focus groups thereby producing items using the language of the subjects interviewed and incorporating the content of qualitative statements made by patients [53]. Developers of future instruments should focus on involving patients from the population of interest. Therefore, creating PROMs which represent patient priorities and items which are acceptable, comprehensive, and relevant to their condition.

Clinicians should be encouraged to use PROMs as evidence suggests the reporting of PROMs, increases patient satisfaction with treatment, and improves adherence to regimes [54,55]. However, using PROMs beyond their intended use may result in data lacking responsiveness and clinical meaning. Instruments in their current format may be inappropriate and contain irrelevant items to the population being studied. This review demonstrated that majority of instruments lacked content validity assessment supporting their use in a CPP population. We identified a single content validity study assessing the SF 36 in a CPP population [49]. Stones et al. identified that questions such as those describing pain did not reflect the episodic/intermittent nature of CPP but rather implied that pain is constant. In addition, participants with CPP found the timeframe of the questionnaire problematic and questions regarding avoidance behaviour with respect to activity and use of analgesia which affected their current pain experience.

It is essential that instruments used by clinicians and researchers are ‘‘fit for purpose’’. This requires adaptation or modification of existing PROMS where the development and subsequent validation of the instrument occurred in a different population compared to that of the study population [56]. Cognitive interviews can be used to adapt existing PROMs by modifying instructions or items in response to patient feedback received and therefore minimizing missing or inaccurate data. Cognitive interviewing serves two purposes: (1) does the instrument content represents the most important aspect of the construct of interest and (2) do respondents

Table 3. Evidence synthesis on content validity for QoL instruments for female CPP

PROM	Responsiveness rating[a]; quality[b]	Comprehensiveness rating[a]; quality[b]	Comprehensibility rating[a]; quality[b]
BDI 2.0	?; very low	−; very low	−; very low
BPI	±; very low	?; very low	?; very low
EHP 30	+; low	+; low	?; very low
EHP 5	+; low	?; very low	?; very low
EQ-5D 3L	?; very low	−; very low	?; very low
EQ-5D 5L	?; very low	?; very low	?; very low
FABQ	±, very low	−; very low	?; very low
HADS	?; very low	?; very low	?; very low
IIP-64	±; very low	?; very low	?; very low
IIP-32	±; very low	?; very low	?; very low
MPI	+; low	?; very low	?; very low
ODI 1.0	?; very low	−; very low	−; very low
ODI 2.1a	?; very low	−; very low	−; very low
PBPQ	?; very low	−; very low	−; very low
SAQ	?; very low	?; very low	−; very low
SF 36 [49]	?; low	?; low	?; low
SF 12	±; very low	?; very low	?; very low
WHOQoL	?; very low	?; very low	?; very low
WHOQoL-Bref	±; very low	?; very low	?; very low

Abbreviations: BDI, becks depression inventory; PBPQ, pain beliefs and perception questionnaire; FABQ, fears avoidance beliefs questionnaire; SAQ, sexual activity questionnaire; BPI, brief pain inventory; HADS, hospital anxiety and depression scale; SF 36, short form survey 36; SF 12, short form survey 12; EHP 30, endometriosis health profile 30; EHP-5, endometriosis health profile 5; IIP 64, inventory of interpersonal problems 64; IIP 32, inventory of interpersonal problems 32; ODI 1.0, oswestry disability index 1.0; ODI 2.1a, oswestry disability index 2.1a; MPI, multidimensional pain inventory.

[a] Ratings can be (+) sufficient, (−) insufficient, (±) inconsistent, and (?) indeterminate.

[b] Quality can be high, moderate, low, and very low.

understand how to complete the instrument including a clear understanding of instructions, interpretation of items, appropriate recall periods, how to use scales, and any other factors that may influence participant responses. Without prior testing of the questionnaire, it is unknown whether patients will encounter difficulties when completing the questionnaire.

Our findings confirm previous reports of poor reporting of qualitative methods with respect to establishing content validity [57]. It is essential that processes for evaluating content validity are transparent and well documented for scientific and regulatory purposes [56]. Multiple guidelines have provided recommendations for demonstrating content validity during PROM development including a literature review, conducting concept elicitation reviews or focus groups, data analysis, item generation, and performing cognitive interviews [53,58−60]. Variations in reporting of a qualitative methodology may be attributed to inadequate dissemination of guidance and knowledge among researchers or a lack of demand from editorial boards of journals [57].

This review was conducted as part of a wider project to establish COS in CPP. COS have the potential to reduce the inconsistencies in outcome reporting. However, the adoption and usefulness of COS may be limited by a lack of recommendations on how to measure core outcomes [61]. There is substantial variation in the methods used to identify, appraise, and select PROMs for COS across a range of medical specialities [61]. Our previous systematic review demonstrated an inconsistent use of PROMs and variation of outcomes reporting in therapeutic trials of women with CPP [4]. Resulting differences in outcome domains, terminology, subscales, and scoring prevent comparison and synthesis of data. The COSMIN/COMET guideline provides a practical four-step method to guide COS developers undertaking this process, including recommendations concerning the selection of measures for a COS: (1) select one instrument per outcome; (2) ensure there is high-quality evidence for content validity, internal consistency, and feasibility of the instrument; and (3) obtain consensus on the instrument [7]. Uptake of the COSMIN guidance will ensure core outcomes are operationalized and consistently measured.

We evaluated content validity of PROMs reporting QoL outcomes in women with CPP. It is the first measurement property to consider when selecting a PROM. Our research group is in the process of evaluating further psychometric properties. These findings will inform future discussions thereby facilitating a consensus because valid and reliable instruments are recommended for the assessment of core outcomes such as QoL.

6. Conclusion

This systematic review has shown poor quality evidence for content validity of PROMs measuring QoL in women with CPP. Developers of future instruments should pay attention to the judicious documentation of qualitative research methods and consider the COSMIN criteria when developing PROMs.

CRediT authorship contribution statement

Vishalli Ghai: Study conception, design, data collection, analysis, and drafting the manuscript. **Venkatesh Subramanian:** Data collection, analysis, and review of draft manuscript. **Haider Jan:** Review of study design and review of draft manuscript. **Stergios K. Doumouchtsis:** Study conception, review of study design, data collection, and draft manuscript.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2022.04.016.

References

[1] Mathias SD, Kuppermann M, Liberman RF, Lipschutz RC, Steege JF. Chronic pelvic pain: prevalence, health-related quality of life, and economic correlates. Obstet Gynecol 1996;87(3):321−7.
[2] Chen I, Thavorn K, Shen M, Goddard Y, Yong P, MacRae GS, et al. Hospital-associated costs of chronic pelvic pain in Canada: a population-based descriptive study. J Obstet Gynaecol Can 2017;39(3):174−80.
[3] Chiarotto A, Terwee CB, Kamper SJ, Boers M, Ostelo RW. Evidence on the measurement properties of health-related quality of life instruments is largely missing in patients with low back pain: a systematic review. J Clin Epidemiol 2018;102:23−37.
[4] Ghai V, Subramanian V, Jan H, Pergialiotis V, Thakar R, Doumouchtsis SK. A systematic review on reported outcomes and outcome measures in female idiopathic chronic pelvic pain for the development of a core outcome set. BJOG 2020;128:1471−6412.
[5] Ghai V, Subramanian V, Jan H, Thakar R, Doumouchtsis SK. CHORUS: an International Collaboration for Harmonising Outcomes, Research, and Standards in Urogynaecology and Women's Health. A meta-synthesis of qualitative literature on female chronic pelvic pain for the development of a core outcome set: a systematic review. Int Urogynecol J 2021;32(5):1187−94.
[6] Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol 2010;63:737−45.
[7] Prinsen CAC, Vohra S, Rose MR, Boers M, Tugwell P, Clarke M, et al. How to select outcome measurement instruments for outcomes included in a "Core Outcome Set" - a practical guideline. Trials 2016;17(1):449. 10.
[8] Jones GL, Kennedy SH, Jenkinson C. Health-related quality of life measurement in women with common benign gynecologic conditions: a systematic review. Am J Obstet Gynecol 2002;187:501−11.
[9] Bourdel N, Chauvet P, Billone V, Douridas G, Fauconnier A, Gerbaud L, et al. Systematic review of quality of life measures in patients with endometriosis. PLoS One 2019;14(1):e0208464.
[10] Traylor J, Chaudhari A, Tsai S, Milad MP. Patient-reported outcome measures in benign gynecologic surgery: updates and selected tools. Curr Opin Obstet Gynecol 2019;31(4):259−66.
[11] Neelakantan D, Omojole F, Clark TJ, Gupta JK, Khan KS. Quality of life instruments in studies of chronic pelvic pain: a systematic review. J Obstet Gynaecol 2004;24(8):851−8.
[12] Terwee CB, Prinsen CAC, Chiarotto A, Westerman MJ, Patrick DL, Alonso J, et al. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. Qual Life Res 2018;27:1159−70.
[13] Prinsen CAC, Mokkink LB, Bouter LM, Alonso J, Patrick DL, de Vet HCW, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res 2018;27:1147−57.
[14] Mokkink LB, de Vet HCW, Prinsen CAC, Patrick DL, Alonso J, Bouter LM, et al. COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. Qual Life Res 2018;27:1171−9.
[15] Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ 2009;339:b2700.
[16] Terwee CB, Jansma EP, Riphagen II, de Vet HCW. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. Qual Life Res 2009;18(8):1115−23.
[17] Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. Spine 2000;25(24):3186−91.
[18] Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336:924−6.
[19] Beck AT, Ward AT, Mendelson M, Mock J, Erbaugh J. An inventory for measuring depression. Arch Gen Psychiatry 1961;4:561−71.
[20] Beck AT, Steer RA, Brown GK. Manual for the Beck Depression Inventory-II. San Antonio, TX: Psychological Corporation; 1996.
[21] Cleeland CS. The brief pain inventory user guide. Available at https://www.mdanderson.org/documents/Departments-and-Divisions/Symptom-Research/BPI_UserGuide.pdf. Accessed November 12, 2020.
[22] Jones G, Kennedy S, Barnard A, Wong J, Jenkinson C. Development of an endometriosis quality-of-life instrument: the endometriosis health profile-30. Obstet Gynecol 2001;98(2):258−64.
[23] Jones G, Jenkinson C, Kennedy S. Development of the short form endometriosis health profile questionnaire: the EHP-5. Qual Life Res 2004;13:695−704.
[24] EuroQol Group. EuroQol–a new facility for the measurement of health-related quality of life. Health Policy 1990;16(3):199−208.
[25] EuroQol Research Foundation. EQ-5D-3L user guide, Version 6. EuroQol Research Foundation; 2018. Available at https://euroqol.org/publications/user-guides. Accessed November 16, 2021.
[26] EuroQol Research Foundation. EQ-5D-5L user guide, Version 3. 2019. Available at https://euroqol.org/publications/user-guides. Accessed November 16, 2021.
[27] Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res 2011;20:1727−36.
[28] Waddell G, Newton M, Henderson I, Somerville D, Main C. A Fear-Avoidance Beliefs Questionnaire (FABQ) and the role of fear-avoidance beliefs in chronic low back pain and disability. Pain 1993;52(2):157−68.
[29] Zigmond AS, Snaith RP. The hospital anxiety and depression scale. Acta Psychiatr Scand 1983;67(6):361−70.
[30] Barkham M, Hardy GE, Startup M. The structure, validity and clinical relevance of the inventory of interpersonal problems. Br J Med Psychol 1994;67(2):171−85.

[31] Horowitz LM, Alden LE, Wiggins JS, Pincus AL. Inventory of interpersonal problems manual. Menlo Park, CA: Mind Garden Inc.; 2003.
[32] Barkham M, Hardy GE, Startup M. The IIP-32: a short version of the inventory of interpersonal problems. Br J Clin Psychol 1996;35(1):21−35.
[33] Kerns RD, Turk DC, Rudy TE. The west haven-yale multidimensional pain inventory (WHYMPI). Pain 1985;23(4):345−56.
[34] Fairbank JC, Couper J, Davies JB, O'Brien JP. The Oswestry low back pain disability questionnaire. Physiotherapy 1980;66(8):271−3.
[35] Fairbank JC, Pynsent PB. The oswestry disability index. Spine (Phila Pa 1976) 2000;25(22):2940−52. discussion 2952.
[36] Williams DA, Thorn BE. An empirical assessment of pain beliefs. Pain 1989;36:351−8.
[37] Williams DA, Robinson ME, Geisser ME. Pain beliefs: assessment and utility. Pain 1994;59(1):71−8.
[38] Thirlaway K, Fallowfield L, Cuzick J. The Sexual Activity Questionnaire: a measure of women's sexual functioning. Qual Life Res 1996;5:81−90.
[39] Ware J, Snow KK, Kosinski MA, Gandek B. SF-36 health survey manual and interpretation guide. Boston, MA: The Health Institute, New England Medical Centre; 1993.
[40] Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992;30:473−83.
[41] Ware JE, Kosinski M, Keller SD. SF-12: how to score the SF-12 physical and mental health summary scales. 2nd ed. Boston, MA: The Health Institute, New England Medical Centre; 1995.
[42] World Health Organization. Programme on mental health: WHOQoL user manual. Geneva: World Health Organization; 1998.
[43] The World Health Organization Quality of Life Assessment (WHOQOL). Development and general psychometric properties. Soc Sci Med 1998;46:1569−85.
[44] Development of the world health organization WHOQOL-BREF quality of life assessment. The WHOQOL group. Psychol Med 1998;28:551−8.
[45] Chiarotto A, Ostelo RW, Boers M, Terwee CB. A systematic review highlights the need to investigate the content validity of patient-reported outcome measures for physical functioning in patients with low back pain. J Clin Epidemiol 2018;95:73−93.
[46] Craxford S, Deacon C, Myint Y, Ollivere B. Assessing outcome measures used after rib fracture: a COSMIN systematic review. Injury 2019;50:1816−25.
[47] Alden LE, Wiggins JS, Pincus AL. Construction of circumplex scales for the inventory of interpersonal problems. J Pers Assess 1990;55(3−4):521−36.
[48] Horowitz LM, Rosenberg SE, Baer BA, Ureño G, Villaseñor VS. Inventory of interpersonal problems: psychometric properties and clinical applications. J Consult Clin Psychol 1988;56(6):885−92.
[49] Stones RW, Selfe SA, Fransman S, Horn SA. Psychosocial and economic impact of chronic pelvic pain. Baillieres Best Pract Res Clin Obstet Gynaecol 2000;14(3):415−31.
[50] Davies CF, Macefield R, Avery K, Blazeby JM, Potter S. Patient-reported outcome measures for post-mastectomy breast reconstruction: a systematic review of development and measurement properties. Ann Surg Oncol 2021;28(1):386−404.
[51] Comins J, Siersma V, Couppe C, Svensson RB, Johansen F, Malmgaard-Clausen NM, et al. Assessment of content validity and psychometric properties of VISA-A for Achilles tendinopathy. PLoS One 2021;16(3):e0247152.
[52] European Medicines Agency Committee for Medicinal Products for Human Use (CHMP). Appendix 2 to the guideline on the evaluation of anticancer medicinal products in man. The use of patient-reported outcome (pro) measures in oncology studies. 2016. Available at https://www.ema.europa.eu/en/documents/other/appendix-2-guideline-evaluation-anticancer-medicinal-products-man_en.pdf. Accessed December 22, 2021.
[53] Brod M, Tesler LE, Christensen TL. Qualitative research and content validity: developing best practices based on science and experience. Qual Life Res 2009;18(9):1263−78.
[54] Chen J, Ou L, Hollis SJ. A systematic review of the impact of routine collection of patient reported outcome measures on patients, providers and health organisations in an oncologic setting. BMC Health Serv Res 2013;13(1):211−24.
[55] Wartolowska K. The nocebo effect as a source of bias in the assessment of treatment effects. F1000Res 2019;8(5):5.
[56] Patient-Reported Outcomes and Medical Devices - Guidance. 2020. Available at https://www.fda.gov/regulatory-information/search-fda-guidance-documents/principles-selecting-developing-modifying-and-adapting-patient-reported-outcome-instruments-use. Accessed December 22, 2021.
[57] Ricci L, Lanfranchi J-B, Lemetayer F, Rotonda C, Guillemin F, Coste J, et al. Qualitative methods used to generate questionnaire items: a systematic review. Qual Health Res 2019;29(1):149−56.
[58] Lasch KE, Marquis P, Vigneux M, Abetz L, Arnould B, Bayliss M, et al. PRO development: rigorous qualitative research as the crucial foundation. Qual Life Res 2010;19:1087−96.
[59] Patrick DL, Burke LB, Gwaltney CJ, Leidy NK, Martin ML, Molsen E, et al. Content validity–establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO good research practices task force report: part 1–eliciting concepts for a new PRO instrument. Value Health 2011;14(8):967−77.
[60] Patrick DL, Burke LB, Gwaltney CJ, Leidy NK, Martin ML, Molsen E, et al. Content validity–establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO Good Research Practices Task Force report: part 2–assessing respondent understanding. Value Health 2011;14(8):978−88.
[61] Ramsey I, Eckert M, Hutchinson AD, Marker J, Corsini N. Core outcome sets in cancer and their approaches to identifying and selecting patient-reported outcome measures: a systematic review. J Patient Rep Outcomes 2020;4(1):77. 12.