Methodological considerations in cross-sectional studies validating the International Classification of Functioning, Disability and Health (ICF) Core Sets: A systematic review.

M. Rojas[a,b], G. Guilera[a,b], E. Arias-Patiño[a,c], M. Barrios[a,b] and J. Gómez-Benito[a,b]

[a] Department of Social Psychology and Quantitative Psychology, Faculty of Psychology, University of Barcelona, Barcelona, Spain; [b] Group on Measurement Invariance and Analysis of Change (GEIMAC), Institute of Neurosciences, University of Barcelona, Barcelona, Spain; [c] Group on Methods and Instruments for Research in Behavioral Sciences, Universidad Nacional de Colombia, Bogotá, Colombia

ABSTRACT

Background: Cross-sectional studies are commonly used to validate International Classification of Functioning, Disability, and Health Core Sets (ICF-CSs). However, the validation quality depends on the methodological approach employed. To date, there is no systematic review of the methodological characteristics and quality of cross-sectional studies validating ICF-CSs. **Objective:** To describe and analyze the methodological characteristics and quality of cross-sectional studies that have validated an ICF-CS. **Methods:** We identified empirical studies published in English and which validated any ICF-CS using a cross-sectional design. The following databases were searched [18-22 November 2022, with an update conducted 7-11 October 2023]: Web of Science (WoS), Scopus, CINAHL, PubMed, Embase, and PsycINFO. Two independent reviewers coded studies that met the inclusion criteria and assessed their methodological quality and risk of bias using the AXIS tool. Synthesis was performed by calculating frequencies and percentages. **Results:** We analyzed 87 articles validating a total of 23 ICF-CSs. In terms of quality, most articles had strengths in terms of consistency between the reported study objectives and the outcome variables measured. However, a large majority do not report the sample size calculation (up to 94.2% in the case of Delphi studies), and few validation studies have been conducted in the WHO regions of Africa and the Eastern Mediterranean. **Discussion:** The quality of cross-sectional studies validating ICF-CSs is satisfactory, although several articles do not describe aspects such as sample size calculation. Validity evidence for ICF-CS studies could be enhanced by conducting more multicenter studies, by replicating ICF-CS validation studies in different WHO regions, and through synthesis of existing research.

***Keywords*:** validation, ICF Core Set, cross-sectional studies, methodological review, report quality.

Implications for Rehabilitation

- Cross-sectional validation studies of ICF-CSs have satisfactory quality, supporting the use of the Core Sets in clinical rehabilitation settings similar to those evaluated here.

- Additional validation studies are required for ICF-CSs that have not yet been validated or for which validity evidence is limited.

- The methodological findings of this review constitute a roadmap that could guide the development and improve the quality of future ICF-CS validation studies.

- Knowing which ICF-CSs are validated through cross-sectional designs is useful for planning and designing interventions and instrument development.

Introduction

Assessing the functioning of a person with a disease is essential for understanding their needs, how their environment affects their functioning, and how the condition impacts their daily life. With this in mind, in 2001 the World Health Organization (WHO) introduced the International Classification of Functioning, Disability, and Health (ICF) as a reference framework for describing and evaluating people's functioning. Based on a biopsychosocial model, the ICF is a universally accepted conceptual model of functioning and disability [1] that is able to describe both the different components of an individual's functioning and their interaction in the specific context in which health care is received [2]. The ICF is composed of more than 1400 categories organized hierarchically, from the most general to the most specific, and it comprises three components*:* (a) *Body functions and structures*, (b) *Activities and participation*, and (c) *Contextual factors*, the latter including both *Environmental* and *Personal factors*. These components interact with a person's health status and determine individual functioning and disability [2]. However, assessing a person's functioning using such an extensive number of categories is impractical within a clinical setting, which is why the ICF Core Sets (ICF-CSs) were created as groupings of categories that sufficiently describe health-related states and which allow both the description and comparison of areas of functioning of a person or group of people with a specific disease or diagnosis [1, 2]. As specific sets of ICF categories for describing a particular health problem or condition, the ICF-CSs can help health professionals to prioritize and focus on the most relevant aspects of an individual's functioning and disability [1, 2]. Currently, 37 ICF-CSs have been developed, covering different chronic and acute health conditions [3]. There are two versions of ICF-CSs: Comprehensive and Brief. Comprehensive CSs include the full set of ICF categories that are relevant to the assessment of functioning for a particular health condition, while Brief CSs consists of the minimum number of categories (from among those included in the Comprehensive CS) that are essential for determining the functioning of an individual with a given health condition. Each type of CS serves a distinct purpose. For instance, suppose it is necessary to create a long-term treatment plan or to have a comprehensive assessment of a person's health and functioning by several specialists. In that case, it is most appropriate to use the Comprehensive CS. Conversely, the Brief CS would be used if a quick assessment or screening of the person's functioning is needed [2].

To develop an ICF-CS, a three-phase guideline has been established to ensure that each CS is (a) based on scientific evidence, (b) reflects the perspective not only of health professionals and experts but also of individuals with a specific health condition, and (c) the health professionals and experts consulted represent a broad range of disciplines of relevance to the health condition [4]. Specifically, the Preparatory Phase involves gathering perspectives from different stakeholders through various methods (i.e., systematic literature reviews, qualitative studies, expert surveys, and empirical studies). Phase I consists of the International ICF Consensus Conference, which leads to the development of the first version of the ICF-CS. Finally, Phase II focuses on the implementation and validation of this initial version in clinical practice or research settings, ensuring the validity of the ICF-CS for assessing and describing people's functioning [4]. The quality of the validity evidence obtained will depend on the methodological approaches and designs used [5].

Phase II validation studies of an ICF-CS seek to accumulate scientific evidence to support the adequacy of the set of categories, demonstrating that they: (a) are easy to understand and use in the clinical context, (b) provide a reliable measure of individual functioning at different points in the health care process, (c) are sensitive and sufficient for describing individuals' functioning and guiding their treatment correctly, and (d) have worldwide applicability, that is, they can be applied to individuals with diverse demographic characteristics and in different cultural contexts without showing changes in their measurement characteristics [4, 5].

There are no pre-established guidelines for validating ICF-CSs [4]. Population characteristics, the variables of interest, and the availability of time and resources all guide the choice of a validation design. Studies may therefore be quantitative or qualitative [6], and the sources of information they use can differ depending on the conditions under which the ICF-CS will be used in practice [2, 7].

Wiegand et al. [8] reviewed ICF implementation studies and found a lack of representativeness due to limited sample size, lack of strength in the evidence by not including control groups, and flaws in statistical procedures. Consequently, they suggest using multicenter and longitudinal designs to generate solid evidence about implementation of the ICF in rehabilitation. An example of a longitudinal design for validating ICF-CSs is the empirical process described by Grill and Stucki [9], which involved a prospective multicenter cohort study. These authors proposed an algorithm

based on three criteria (truth, discrimination, and feasibility) for selecting ICF-CS categories and validating their implementation in a clinical context with acute and post-acute characteristics. Notwithstanding these developments, cross-sectional designs are commonly used to validate ICF-CSs, and there are no specific guidelines regarding how they should be applied.

Cross-sectional validation designs can be used in those cases where studying the time variable or the effects of exposure to a given treatment is of no relevance. One advantage of this type of design is that information can be collected from individuals, groups, organizations or countries within a relatively short period [10] using quantitative, qualitative or mixed methods, including surveys, one-on-one interviews, panel interviews, and focus groups [11]. Cross-sectional designs usually involve the calculation of frequency distributions and comparisons between subgroups, although associations between variables of interest can also be explored [10].

When conducting a cross-sectional study, it is crucial to have a well-defined sampling frame, appropriate participant selection methods, defined outcomes, and reliable data analysis procedures. The extent to which the findings will be generalizable to the population for which a specific ICF-CS is applicable depends on these factors and the ability to control potential sources of bias [12].

Although a recent scoping review [6] identified the methods used in studies validating ICF-CSs, no in-depth analysis of methodological practices or the quality of existing reports has yet been published. The authors of the scoping review classified articles according to the methods used in the validation, namely qualitative, quantitative or mixed, but they did not specify whether these correspond to cross-sectional or longitudinal designs. Neither did they examine in detail their methodological characteristics or quality. They do, however, note that the studies differed in quality attributes, highlighting the need to evaluate this aspect in future reviews. The aim of the present review is therefore to describe and analyze the methodological characteristics and quality of cross-sectional studies validating ICF-CSs so as to inform current practice when conducting such studies.

Method

This systematic review followed the procedures described in the PRISMA 2020 statement [13].

Eligibility criteria

Included articles had to meet the following criteria: (a) report an empirical validation study of an ICF-CS, (b) validate the ICF-CS in either its Comprehensive or Brief version, (c) use a cross-sectional design, and (d) be available in full text in English. Editorials, conference presentations, books and book chapters, and literature reviews were excluded. Also excluded were studies not directly validating an ICF-CS (e.g., instruments derived from ICF-CS categories), longitudinal and cohort studies, secondary data analyses, and qualitative studies that did not involve data quantification (e.g., narrative analysis).

Search strategy

To identify relevant articles, we searched a variety of health and social science databases: Web of Science (WoS), Scopus, Cumulative Index to Nursing and Allied Health Literature (CINAHL), PubMed, Excerpta Medica Database (Embase), and PsycINFO. Three blocks of keywords were used for the electronic search. The first block contained keywords related to ICF-CS validation studies, referring to Phase II of the development of a CS (e.g., "valid*"). The second block referred to the ICF system (e.g., "International Classification of Functioning, Disability and Health"), and the last block contained keywords related to the CS (e.g., "core-set"). Within each block, the keywords were combined using the Boolean operator OR, while the blocks themselves were combined using AND. A detailed search syntax for each database can be found in the Supplementary Material, Appendix A. The search was restricted to articles published from 2001, when the ICF classification was created, until November 2022 (with an update of the search performed in October 2023).

Screening and selection of studies

Search results were loaded into an Excel database to eliminate duplicate references. Two reviewers (MR and EA) independently examined all titles and abstracts to obtain a preliminary selection of articles based on the inclusion and exclusion criteria. Any disagreements were resolved through consensus. The reference list of the recently

published scoping review [6] was also screened to identify any other records that might be eligible for inclusion.

Data extraction and coding

Articles that met the inclusion criteria were coded independently by two reviewers (MR and EA). The descriptive variables that were coded included the sample source, the type of ICF-CS (Comprehensive vs. Brief), the target disease or condition, and sample size, among others. Variables referring to the methodological characteristics of the cross-sectional studies, such as procedures used to analyze the information and cut-off points for selecting the ICF categories, were also recorded.

We conducted a two-step coding process to analyze the methodological limitations reported in the studies. In the first step, each coder independently extracted the phrases referring to methodological limitations and grouped them according to the frequency with which the themes repeated. In the second step, the coders analyzed and discussed the thematic groupings made in the first step of analysis and reached a consensus regarding the general categories of the group.

The quality of articles was also evaluated using the Appraisal tool for Cross-Sectional Studies [14] (AXIS tool). The AXIS tool allows for a critical approach to the study design and quality of reporting, as well as the risk of bias in cross-sectional studies. It consists of 20 elements that qualitatively evaluate each aspect of a study design. The two coders performed this quality assessment independently for all selected articles, and agreement between them was assessed by calculating Cohen's kappa statistic in SPSS. Coefficients were interpreted as follows: 0.4 – 0.6, fair agreement; 0.6 – 0.75, good; 0.75 and above, excellent [11].

Results

1. Literature search results

A total of 1478 articles were initially retrieved by the search. After eliminating duplicates, a total of 545 records were obtained. The abstracts of these 545 records were reviewed, and 104 papers were selected. To confirm which of these articles should definitely be included, we obtained the full text of each publication. Full text review led us to retain 86 studies and exclude 18 that did not meet the inclusion criteria. Screening of the reference list of the recent scoping review [6] yielded an additional record, and

thus the total number of articles included was 87. For a more detailed description of the record selection process, see the PRISMA flow chart in Figure 1.

2. *Research techniques used to validate ICF-CSs*

Four kinds of studies were identified based on the research technique used to validate an ICF-CS (i.e., questionnaire, individual interview, focus groups, and Delphi method). In questionnaire studies, a list of the ICF-CS categories to be evaluated was presented to participants in the form of a questionnaire, enabling them to rate the relevance of each category to people with the health condition in question (hereinafter, ICF-CS questionnaire studies). Ratings are given using a Likert-type scale, from 0 (not a problem, not relevant) to 4 (core problem, highly relevant). Of these 33 studies, 84.9% (n = 28) used an additional instrument specific to the disease or condition or a general tool assessing health and functioning, such as the 36-Item Short Form Health Survey (SF-36).

In a second type of study, the researchers conducted individual interviews to inquire about people's functioning and relate it to the corresponding ICF-CS. For example, individuals with different health conditions are asked open-ended questions about their functioning with respect to different components of the ICF (e.g., "If you think about your body and mind, what does not work the way it is supposed to?"). The information obtained in these interviews is then subjected to content analysis using the meaning condensation procedure [15, 16]. This technique divides the data into units of meaning consisting of a few words or phrases with a common theme. The concepts underlying these units are then identified and a linking process is carried out, which consists of matching the concepts identified in the content analysis with the ICF categories with which they are most related [4, 17].

Other studies used focus groups as a technique. In these studies, which involve small groups of individuals, a moderator poses open-ended questions about problems represented by the ICF components. These questions are similar to those asked in the interview studies and they allow people to express their opinions and discuss them within the group. The information obtained in the focus group sessions is analyzed using the meaning condensation procedure to link the identified concepts with the ICF categories, similar to the interview studies.

Finally, some studies used the Delphi method to validate an ICF-CS. This technique typically involves recruiting experts in the health condition to which the CS being studied refers, and then, in successive rounds, requesting their opinions on the problems they believe most affect peoples' functioning, as well as treatment goals. In the first round, questionnaires are distributed to gather information about the functional problems of individuals with the health condition. This questionnaire might consist of open-ended questions such as: "What are the patient problems, patient resources, and aspects of the environment treated by [name of profession] in people with [health condition]?". Alternatively, the questionnaire may ask about the relevance or pertinence of each category of the ICF-CS to the functioning of people with the health condition in question. The responses are then linked to the ICF categories and presented again for further input regarding the impact on functioning. In the final round (usually the third), the experts are presented with the ICF categories and feedback on what percentage of experts considered each ICF category a problem, the outcome of which is a final expert rating. The ICF categories resulting from this process are then compared with the corresponding ICF-CS.

INSERT FIGURE 1 HERE: PRISMA flow chart.

3. General characteristics of the studies

The 87 studies examined the ICF-CSs of 23 diseases or health conditions. The two most frequently validated CSs were those for stroke (n = 12; 13.8%) and low back pain (n = 9; 10.3%). Regarding the version that was validated, 81.6% of studies (n = 71) validated the Comprehensive CS, 12.6% (n = 11) the Brief CS, and 5.8% (n = 5) both. Concerning the type of study, it was found that 33 studies used questionnaires as the data collection technique (37.9% of the total of 87 studies), 8 studies involved interviews (9.2 %), 10 studies used focus groups (11.5%), and 36 employed the Delphi method (41.4%). For further details regarding the characteristics of the 87 studies, see Table 1.

Among ICF-CS questionnaire studies, 97% (n = 32) targeted individuals with particular health conditions, while the remaining 3% (n = 1) explored the perspective of health professionals [18]. Regarding interview studies, 87.5% (n = 7) focused on persons with specific health conditions, while 12% (n = 1) explored the clinician perspective [19]. Focus groups exclusively involved participants with specific CS-related conditions,

whereas all Delphi studies engaged experts defined by their experience in treating specific diseases.

INSERT TABLE 1 HERE: Characteristics of the cross-sectional validation studies of ICF-CSs included in this review.

Cross-sectional ICF-CS validation studies have been conducted across all WHO regions. Europe is the most widely represented region (in 70 studies, 80.5%), while the Eastern Mediterranean and Africa are the least represented (in 27 studies each, 31%). Studies using an ICF-CS questionnaire and the Delphi method have been carried out in all regions, but interview studies have only been conducted in three regions (i.e., Europe, Americas, and Western Pacific) and focus groups in two (i.e., Europe and Western Pacific) (see Table 2). Although ICF-CS validation studies involving the Delphi technique are the most widely used in each region, five studies (13.9%) included experts from just a single WHO region.

INSERT TABLE 2 HERE: Number of studies by WHO region and the research technique used for validation.

Of the 33 ICF-CS questionnaire studies, 66.7% (n = 22) were multicenter, with 36.4% (n = 8) conducted in more than one country. Among the eight interview studies, 50% (n = 4) involved multiple centers, though none spanned more than one country. Similarly, 60% (n = 6) of the focus group studies were multicenter, but all were confined to a single country.

In studies using the Delphi technique, experts were sought from various sources, including research centers, hospitals, universities, and through online scientific databases. Consequently, the exact number of participating centers is unknown. Of the 36 studies of this kind, 88.9% (n = 32) were conducted in multiple countries, with only 11.1% (n = 4) being limited to a single country.

4. Methodological characteristics of studies

The methodological characteristics we analyzed were sample size and type of sampling, the techniques employed to analyze data, and the methodological limitations reported in the studies. We also examined the cut-off (threshold) points established by researchers for deciding which categories merited inclusion in an ICF-CS (or in the case of Delphi

studies, the criterion for defining expert consensus about ICF-CS categories).

4.1. ICF-CS questionnaire studies

In these studies (n = 33) the sample size ranged from 99 to 972 participants. Regarding sampling (see Table 3), in 31 studies (93.9%) it was possible to establish what type of sampling was used, either because the authors reported it (n = 19; 61.3%) or because it could be inferred from the information provided in the study (n = 12; 38.7%). In all cases, sampling was either consecutive or convenience.

INSERT TABLE 3 HERE: Type of sampling used in cross-sectional ICF-CS validation studies.

Regarding the techniques used to analyze the data collected, in most cases (n = 22; 66.7%) frequencies and percentages were used to indicate the prevalence of ICF-CS categories (i.e., how representative they are of the problems experienced by people with the health condition that is the target of the ICF-CS). Several studies also correlated CS categories with scores on condition-specific tests for assessing a person's health (n = 19; 57.6%). Table 4 presents other analytic techniques employed in these studies.

INSERT TABLE 4 HERE: Techniques for analyzing data in cross-sectional ICF-CS validation studies.

One of the most common methods (n = 22; 66.7%) used in ICF-CS questionnaire studies to decide which categories were suitable for assessing functioning was analyzing the relative frequency with which they represented a problem for the study sample (i.e., an indicator of prevalence). Researchers established a cut-off point (i.e., a minimum or maximum percentage of the sample that presents problems in the category) to discard ICF-CS categories considered unsuitable for assessing functioning. Fifteen studies (45.5%) reported a specific cut-off point. These studies established six different cut-off points, as shown in Table 5, with the most commonly used being >10%.

Regarding the limitations of ICF-CS questionnaire studies, these relate to the characteristics of the sample and its representativeness. Eighteen studies (54.6%) mentioned that the sample was limited to a specific country or region, diagnostic group, level of impairment, or sex. Thirteen studies (39.4%) noted the lack of

representativeness of the sample, while ten papers (30.3%) highlighted a possible risk of bias due to a lack of standardization or training in the application and scoring of the CS. Ten articles (30.3%) mentioned possible limitations related to sample size, such as the small sample size potentially affecting the statistical power of the analyses. Five articles (15.2%) highlighted limitations in generalizing results due to having collapsed ICF ratings into a dichotomous variable or to modifications made in the application of ICF categories. Finally, four studies (12.1%) noted problems with the outcomes selected, and three (9.1%) mentioned the use of arbitrary cut-off points (cut-off criteria).

INSERT TABLE 5 HERE: Cut-off (threshold) point for deciding which ICF-CS categories are suitable for assessing functioning, based on the minimum percentage of the target population who experience problems in a given category.

4.2. Interview studies

The sample size in the eight studies that used interviews ranged from 7 to 113 participants. In seven studies (87.5%) it was possible to establish the type of sampling used, either because the authors clearly stated this (n = 4; 57.1%) or because it could be inferred from their description of how the sample was recruited (n = 3; 42.9%). Purposive or convenience sampling was used in five studies (71.4%) and the maximum variation strategy in two (28.6%) [19, 89]. Five studies (62.5%) used the saturation technique as a quality criterion to ensure the exhaustiveness of the information collected. Saturation is defined as the point at which data analysis yields no further information on the topic under investigation [91]. Three different criteria were used to define that saturation had been reached. The strictest criterion, used in one study (20%), defined saturation as being reached if, when linking the data, no additional second-level ICF categories appeared in five successive interviews [26]. Another study (20%) stated that saturation was reached when no new second-level ICF categories were found in three consecutive interviews [90], while three studies (60%) defined saturation as being reached when two consecutive interviews yielded no new second-level ICF categories [19, 78, 89]. The minimum number of interviews required to meet the saturation criterion ranged from seven in one study [19] to a maximum of 35 in another [26].

Researchers also used the linking procedure to link concepts identified in the interviews to ICF categories. In the vast majority of cases (n = 7; 87.5%), linking was combined

with the meaning condensation procedure to analyze the content of interviews. Cohen's kappa was also used in the majority of interview studies (n = 6; 75%) to assess agreement among the researchers who performed the linking procedure. Finally, one study [26] analyzed the relative frequency with which the ICF categories appeared in interviews, the cut-off point for considering an ICF-CS category as valid being that the problem it referred to was mentioned by at least 30% of participants.

Three main areas were identified when considering the limitations of interview studies. The most common limitations were small sample sizes, lack of representativeness across countries, and the non-representativeness of the different levels of the health conditions of the CS being validated (n = 8; 100%). In addition, several limitations related to data collection were identified (n = 7; 87.5%). These limitations included the use of open-ended questions that did not target specific ICF components, thus limiting participants' ability to provide in-depth insights into distinct CS categories. One study (12.5%) also reported challenges related to the profile of clinicians who conducted the interviews, such as their level of training in this respect. Cultural factors also influenced the exploration of specific ICF components, for example, functional problems in categories related to defecation, urination, and sexual functions. Finally, limitations were mentioned in relation to the linking procedure (37.5%; n = 3), primarily concerning potential influence or bias from the coders' professional background. One study did not report any limitations [28].

4.3. Focus group studies

The number of focus groups carried out in these studies ranged from 5 to 15, while the number of group participants ranged from 21 to 60 people. The type of sampling used was identifiable in all ten studies, although in three (30%) it had to be inferred from the authors' description of the method. The sampling method used was the maximum variation strategy in six studies (60%) and purposive sampling in four (40%). Similar to what occurred in interview studies, the saturation technique was used to determine the exhaustiveness of the data obtained. The most stringent criterion required that two consecutive focus groups yielded no new second-level ICF categories in comparison with previous groups (n = 4; 40%). Other less restrictive criteria stated that saturation was achieved when no additional second-level ICF categories emerged in relation to the previous focus group (n = 2; 20%) [62, 70], or when fewer than 5% of the ICF

categories identified were new, either across two successive groups (n = 3; 30%) or in a single additional group (n = 1; 10%) [101].

Regarding the techniques used for analyzing data, all ten focus group studies (100%) performed linking and content analysis. Two approaches to content analysis were described: nine studies (90%) used the meaning condensation procedure, while the other study (10%) [70] used open and axial coding to identify domains in the transcriptions. All ten studies (100%) used Cohen's kappa to assess inter-observer reliability for linking. Table 4 shows additional techniques that were used to a lesser extent. Limitations in focus group studies were mainly related to the participation of people from a single country or region (n = 10; 100%) or with a specific disease subtype (n = 5; 50%). Methodological limitations were also mentioned in relation to focus group procedures (n = 6; 60%), linking methods (n = 6; 60%), the saturation criterion (n = 5; 50%), and clinical sample selection (n = 3; 30%). A limitation affecting the linking procedure is that the specific occupation of health professionals can bias the process. Regarding studies that used the saturation technique, some noted that having reached the pre-established criterion they nevertheless conducted one or more further groups and found that new second-level ICF categories continued to emerge. Open-ended questions, procedural changes, and applications in clinical settings also affected data collection.

4.4. Delphi studies

The Delphi studies had a maximum sample size of 638 experts in one paper [79] and a minimum of 10 participants in another [82] (data correspond to the third Delphi round). In 35 studies it was possible to define the type of sampling used, either because the authors made this explicit (n = 27; 77.1%) or because it could be inferred from their description of how experts were recruited (n = 8; 22.9%). Of these 35 studies, 25 (71.4%) used purposive or convenience sampling, eight (22.9%) used a snowball strategy, and two (5.7%) [21, 98] employed the maximum variation strategy.

Regarding the techniques for analyzing data, linking was used in almost all studies (n = 35; 97.2%) as a strategy to link the concepts that emerged from the Delphi questionnaire applied in round 1 to ICF categories. The reason why linking was not used in the remaining study [43] was that the initial Delphi questionnaire included the corresponding ICF-CS categories and asked experts for their opinion on their relevance

for assessing functioning in the target population. Among studies that did perform linking, 32 (88.9%) calculated Cohen's kappa to determine inter-rater reliability. Additionally, all the Delphi studies (n = 36; 100%) used frequencies and percentages to establish the level of consensus among experts in rounds 2 and 3 of the process regarding the ICF categories that impacted people's functioning. Table 4 provides more information on the data analysis techniques used in Delphi studies.

In the Delphi procedures, a threshold was established to define whether, in the final round, there was consensus among experts that an ICF category was relevant to functioning and, therefore, should form part of the corresponding CS. Most Delphi studies (n = 34; 94.4%) indicated this threshold in the form of a percentage; the remainder (n = 2; 5.6%) did not report this percentage [76, 77]. Three different cut-off points were established to define consensus: 75% of experts (n = 30; 88.2% of studies), 70% (n = 3; 8.8%), and 78% (n = 1; 2.9%) [36]. Several studies (n = 7; 19.4%) also established a threshold during the initial round of the Delphi process, stipulating that only those categories mentioned by at least 5% of participants in the categorization process would be subjected to expert review in subsequent rounds. Other study [43] chose to re-evaluate in subsequent rounds only those categories that attained a predefined level of expert agreement, specifically those falling within the range of 25% to 75% agreement, or those surpassing 30% but under 70%.

Regarding the limitations reported in Delphi studies, 16 articles (44.4%) mentioned the representativeness of the sample, the under-representation of some WHO regions (Africa, Eastern Mediterranean, and South-East Asia), and the over-representation of European countries (n = 6; 16.7%) in general, and Germany in particular. Nine studies (25%) stated that administering the questionnaires in English and online constituted barriers to experts' participation in some regions. On a related issue, six studies (16.7%) highlighted the impossibility of carrying out random sampling of participants. Regarding the cut-off point for reaching consensus, six articles (16.7%) mentioned that 75% agreement is arbitrary, and two studies (5.6%) [43, 23] used different cut-off points between rounds, which may affect the results. With respect to the linking procedure, twelve studies (33.3%) reported limitations related to the reliability and generalizability of results. Five studies (13.9%) achieved low levels of agreement among coders, and a further seven (19.4%) considered that it was not clear whether coders with other professional profiles would have made different decisions.

5. Methodological and reporting quality

Two coders independently assessed the quality of articles using the AXIS tool. Inter-coder reliability was examined by calculating Cohen's kappa for AXIS item ratings as a whole. The coefficient obtained (.89) can be considered excellent.

The results for AXIS items assessing quality and potential bias indicate that the research objective, design, and reference population are well-defined in cross-sectional validation studies using an ICF-CS questionnaire, interviews, focus groups, or Delphi techniques (see items 1, 2, and 4 in Table 6).

As for the justification of sample size (see item 3 in Table 6), this requirement was generally not fulfilled in studies that applied an ICF-CS questionnaire, conducted interviews or used the Delphi technique. Conversely, 90% of focus group studies (n = 9) did justify how the sample size was determined and the method used to do so, which in this case was by means of the saturation technique.

The large majority of studies that used interviews (87.5%; n = 7), focus groups (90%; n = 9) or the Delphi method (97.2%; n = 35) explained how the sample was obtained, what the inclusion and exclusion criteria were, and how participants were selected (see item 6 in Table 6). By contrast, 60.6% (n = 20) of ICF-CS questionnaire studies did not provide complete information about how the sample was obtained and/or about how participants were selected.

In the vast majority of studies, non-responders (see item 7 in Table 6) were not identified, analyzed or described (n = 33, 100% of ICF-CS questionnaire studies; n = 8, 100% of interview studies; n = 10, 100% of focus group studies; n = 30, 85.7% of Delphi studies). None of the studies except those using the Delphi technique provided information about the response rate and whether non-response represented a risk of bias when interpreting the results (see item 13 in Table 6). Among Delphi studies, five (14.3%) identified a risk of non-response bias, and of these, only one reported information on non-respondents [51]. A further and related issue is that many studies did not report the characteristics of participants who did not respond (item 14 in Table 6). This was the case for all the ICF-CS questionnaire, interview, and focus group studies, and for 14.3% of Delphi studies.

Regarding the appropriateness of outcome variables (see item 8 in Table 6), most of the studies that used an ICF-CS questionnaire (75.8%), interviews (75%) or the Delphi method (100%) measured outcome variables that were appropriate to the study aims. However, four studies that used focus groups (40%) and eight ICF-CS questionnaire studies (24.2%) evidenced a risk of bias in this variable. This was due, for example, to the use of indices or instruments that were not relevant to the population, or the inadequacy of the questions used for exploring all components of the ICF-CS (e.g., open-ended questions). As regards item 9 of the AXIS tool, which refers to measuring outcome variables using previously tested or piloted measures and instruments, only two Delphi studies (5.6%) provided information on whether the Delphi questionnaire had been piloted [36, 63].

Overall, most of the studies reviewed meet the criteria established for methodological and reporting quality (see Table 6). For example, most studies showed coherence in the presentation of results across tables and written sections of the article (see item 15 in Table 6). Almost all studies (93.9% of ICF-CS questionnaire studies, and 100% of those using interviews, focus groups or the Delphi method) presented results for the analyses described in the methodology. In addition, in all but three of the articles reviewed (96.6%) the results justified what was stated in the discussion and conclusions (see item 17 in Table 6). Furthermore, most studies (97.7%; n = 85) acknowledged the limitations of their respective research (item 18 in Table 6).

One notable finding from application of the AXIS tool relates to whether the studies obtained ethical approval or informed consent. It can be seen in Table 6 (item 20) that this was obtained by all studies involving use of an ICF-CS questionnaire or interviews, and by 90% of focus group studies. By contrast, half of the Delphi studies (n = 18) did not provide any information on this issue, while two papers (5.6%) [21, 102] stated that it was unnecessary to seek ethical approval due to the nature of the study, although the research did adhere to the ethical standards of the Declaration of Helsinki.

INSERT TABLE 6 HERE: Results obtained when applying the AXIS tool to analyze the quality of cross-sectional studies validating ICF-CSs.

Discussion and Conclusions

The present systematic review offers a comprehensive overview of cross-sectional studies that validate at least one of the components of an ICF-CS. Our review examined 87 articles that validated 23 ICF-CSs, focusing on their methodological characteristics and quality. Stroke and low back pain were the most frequently studied health conditions, with the majority of studies using either the Delphi method or an ICF-CS questionnaire. Study participants were predominantly individuals with a given health condition, and nearly half of the studies were conducted in a single country or center. Cross-sectional validation studies have yet to be conducted for 16 ICF-CSs (see Appendix B, Supplementary Material), and notably, none of the studies reviewed included caregivers.

Although most studies validated the Comprehensive version of the CS, 12.6% focused on the Brief version. This is an interesting finding, given that all the ICF categories of the Brief version are included in the Comprehensive CS. A possible explanation for why some studies opted to validate the Brief version is that the shorter CS may be of greater interest to some researchers because of its clinical and practical implications, that is to say, it considers the essential categories for describing functioning in a given health condition, and its simplicity and shorter administration time make it more suitable for use in the clinical setting [2]. Five studies validated one or two components of an ICF-CS [25, 31, 58, 97, 100], which may be explained by the clinical interest of having indicators related to treatment goals. For example, the Activity and Participation component categories represent the most critical daily limitations and restrictions that people experience [31].

It is worth noting that although the general characteristics of the studies in our systematic review align with those in the scoping review conducted by Karlsson et al. [6] (e.g., the predominance of ICF-CS validation studies using the Delphi method and the majority of validation papers focusing on stroke and low back pain), there is a concerning lack of progress in certain critical aspects of ICF-CS validation. For instance, from 2019 (last year of the scoping review by Karlsson et al. [6]) to 2023 (last year of our systematic review), we found only three additional CSs validated with cross-sectional methodology (traumatic brain injury, musculoskeletal conditions in acute care and musculoskeletal conditions in post-acute care). This is an indication that there has been insufficient effort to validate additional CSs, at least through cross-sectional

validation studies. Expanding the range of validated CSs would provide health professionals with more comprehensive tools to assess the functioning of individuals with a broader spectrum of diseases or clinical conditions within the ICF framework.

ICF-CS questionnaire studies

Studies employing an ICF-CS questionnaire aim to validate the functional level of people with a given health condition, and thus they require a sample that adequately represents both the condition under investigation and diverse WHO regions. However, our analysis reveals that almost half of the studies used samples corresponding to single countries, regions, diagnostic categories, stages of impairment, or genders. This limited scope hinders the broader applicability of the results obtained. Only a modest 36% of these studies were the result of international multicenter research efforts.

Sampling frame selection in cross-sectional studies hinges on the research objectives and the desired level of representativeness. Here we found that the choice of sample often depends on researchers' access to individuals with a specific health condition, which typically leads to hospitals being the primary data collection site. However, this can become a limitation due to selection bias, insofar as the sample may predominantly include in-patients or outpatients, who by definition will likely have different levels of impairment in functioning, and hence the results obtained will not be generalizable to the whole population with the health condition being studied [24, 48].

Regarding the cut-off points used for validating ICF-CS categories, we found that a variety of thresholds were employed, often without clear supporting criteria. The procedure for constructing an ICF-CS suggests that a category can be considered valid if it is reported as problematic by at least 20% of persons with the health condition. However, it is acknowledged that the exact percentage depends on the data collected [4]. This variability in cut-off points potentially influences the number of validated categories. Sensitivity analyses or simulation studies are therefore warranted to determine the need for validation cut-off points and to identify optimal cut-off values based on sample size and the number of categories.

In the cross-sectional validation studies of ICF-CSs that we reviewed, frequencies and percentages were commonly used to provide descriptive information about the prevalence of different functional problems in a given sample. Some researchers also considered it of interest to explore the relationship between the categories of an ICF-CS

and clinical test scores or other variables through correlational analyses, and in some cases, this involved examining mean differences between subgroups. Although some studies also sought to analyze the dimensionality of an ICF-CS, only six studies [33, 50, 52, 53, 67, 95] examined score invariance by analyzing differential item functioning (DIF).

The absence of sample size calculations can lead to bias and compromise results due to inadequate statistical power or imprecise estimates [104]. For example, the sample size was problematic in some studies that used psychometric IRT (item response theory) modeling and regression statistics [33, 50, 53, 67]. Striking a balance is crucial, however, as overly large samples can impact model fit, while small samples may limit statistical power to detect significant effects. Factors such as pilot studies and low-prevalence clinical populations contribute to the omission of sample size calculations, while studies from multicenter endeavors might omit such details for brevity. Applying established criteria for sample sizes, such as Hackshaw's proposal [12], and using straightforward calculation methods, as in Boateng's recommendation of ten observations per item or 200 to 300 total observations, can guide good practices in validation studies [105]. Effect size calculation software can also aid in this regard [104].

Concerning rating/scoring procedures, some studies acknowledged the potential bias introduced by the lack of operationalization of ICF categories [18, 27, 33, 52, 72, 73, 91]. This challenge stems from the complexity of describing certain individual, clinical, and etiological factors using predefined categories, as well as the absence of standardized instructions for scoring ICF-CS categories, which means that practitioners may understand the categories differently.

Previous research [106, 107] suggests that the rater's profession may be a source of DIF, which could impact the scoring of ICF categories. Importantly, this DIF persists even when professionals are trained in how to use and score an ICF-CS [107]. However, systematic research on this topic remains limited. Response process validity studies would offer insights into professionals' ICF-CS scoring practices, shedding light on the cognitive processes governing their understanding and categorization of CS categories. This aspect is crucial for evaluating the appropriateness of the processes employed in defining functioning [7, 108, 109].

Interview and focus group studies

Validation studies that employed qualitative methodologies such as focus groups and interviews mainly explored the perspective of individuals with a specific health condition. One study of this kind examined the perspective of practitioners [19]. Just over half (55%) of the interview and FG studies were multicenter, although in each case, centers were located in a single country. Furthermore, these studies were predominantly conducted in European countries. Beyond seeking the generalizability of their results, these studies primarily aim to identify the most important aspects for individuals in terms of functioning, environment, and personal factors [4], and to examine to what extent these aspects are captured by the respective ICF-CS.

According to Vasileiou et al. [110], studies involving interviews and focus groups often fail to adequately justify sample size, being guided instead by pragmatic criteria such as the volume of data to be analyzed or the influence of disciplinary or publication standards. However, this was the case for only a minority of the studies reviewed here, predominantly interview studies. Those studies that did justify their sample size explained that it was achieved mainly by data saturation.

In cross-sectional studies involving qualitative data, the methods used to determine the required sample size are crucial [11, 111]. Bryman [11] proposed five criteria for determining sample size: a) saturation, b) minimum requirements for an adequate sample, c) theoretical underpinnings of the research, d) the heterogeneity of the population from which the sample is drawn, and e) the research question. For their part, Hennink et al. [112] identified six parameters influencing saturation: a) purpose, b) type of codes, c) stratification of groups, d) the number of groups per stratum, and e) type and degree of saturation. In the present review, we found that a stop criterion for data saturation was commonly used, and sample stratification was absent in most studies; only one study evidenced the theoretical framework from which data were analyzed [70]. While our findings indicate similar saturation ranges to those obtained using empirical saturation criteria in other interview and focus group studies [113], almost half of the studies considered this criterion unreliable [16, 17, 35, 58, 89, 94, 101]. Some studies even included additional data to verify whether new categories emerged after reaching saturation [17, 35]. Interestingly, one study regarded this criterion as a philosophical concept rather than a scientific cut-off [35]. These findings highlight the epistemological dimensions of qualitative research and how researchers interpret the

notion of saturation within their specific theoretical frameworks. Vasileiou et al. [110] concluded that sample size justification in qualitative studies published in health scientific journals was often inaccurate due to generalizability being conceived in nomothetic terms. This could lead some authors to consider small samples as insufficient, without justification.

To comprehensively address these concerns, we recommend that researchers enhance the reporting of saturation criteria by considering various factors of influence, especially those intrinsic to the study. It is essential to expand the descriptive information of participants, incorporating relevant variables such as sex, age, diagnosis, disease severity, and other factors pertinent to the health condition under investigation. Furthermore, it is crucial to provide a clear description of the theoretical framework used for data analysis and to justify the chosen sampling method [111]. In addition to the widely recognized stopping criterion introduced by Coenen et al. [16], researchers could consider implementing supplementary strategies, such as using the number of strata and the number of groups per stratum as critical criteria to identify an adequate sample size for achieving significant saturation. Another useful strategy involves randomizing the order in which data are analyzed, as this sequencing may potentially impact saturation [113]. By integrating these complementary approaches, researchers can enhance the rigor and validity of their qualitative studies.

Regarding the use of open-ended questions in ICF-CS validation studies, several reports [17, 26, 70, 90] have highlighted challenges in validating specific components, such as environmental factors or categories related to sensitive topics. The authors noted that using open-ended questions without explicit references to health domains limits the comprehensiveness of collected information. Consequently, participants may avoid addressing sensitive issues, and moderators could inadvertently stimulate unforeseen avenues of thought. Omissions may also occur in persons with some form of cognitive or psychiatric impairment. Prior research has shown that employing semi-structured interviews and focus groups with an ICF-based approach identifies more categories than do open-ended procedures [16, 78, 114]. By integrating ICF chapter names into open-ended questions, researchers encourage participants to discuss the entirety of ICF-CS components, including sensitive issues like toileting or sexual functions.

The data analysis and sampling procedures used in interview and focus group studies were those suggested in the preparatory study guide for developing an ICF-CS. In some

cases, frequencies and percentages were also analyzed. The use of these methods confirms that the primary aim of these studies was to provide evidence of content validity.

Concerns have been raised about use of the linking procedure in ICF-CS validation studies involving interviews and focus groups, due to the potential for coder bias. Although it is considered the gold standard [115], nearly half of the qualitative studies reviewed here noted limitations with the procedure. Thus, while most studies included quality controls such as reliability checks and multiple raters, concerns persist due to the subjective nature of the linking procedure, whereby experts may interpret concepts differently based on their professional background and experiences.

One response to these concerns was proposed by Cieza et al. [116], who suggested the need to refine existing linking rules to improve the transparency of the process and, ultimately, its reliability. They emphasized the importance of documenting the perspective used in the linking process and distinguishing between capacity and performance, which is vital for precise linking and result interpretation.

Delphi studies

Delphi studies have been employed to validate the comprehensive version of ICF-CSs from the perspective of practitioners. These studies typically offer evidence of content validity by identifying the most common presenting problems among persons with a specific health condition. Additionally, they assess and determine the treatment goals of practitioners from various specialties, comparing them with ICF-CS categories to establish the relevance of different categories.

In both the initial studies conducted to develop an ICF-CS and subsequent validation studies, it is essential to adopt a global perspective and include experts from all six regions of the WHO [4]. The Delphi studies reviewed here showed an overrepresentation of experts from European countries, with relatively few experts from the African, Eastern Mediterranean, and South-East Asian regions. The authors identified language issues (e.g., questionnaires solely in English) and limited internet access as potential reasons for low participation rates in these regions [21, 30, 32, 61, 86, 87, 96, 98]. Interestingly, only seven studies [79, 80, 81, 82, 83, 84, 85] tried to include questionnaires in multiple languages to encourage broader expert engagement. Although this did increase the participation of experts from these regions, the

improvement was not considerable. In addition to the aforementioned issue of internet access, this may reflect the smaller numbers of mental health-related professionals in some countries. Whatever the case, the reasons for low participation from these regions should be studied in more depth in the future.

The use of appropriate sampling techniques is one way of ensuring the representativeness of samples and avoiding selection bias. Many Delphi studies use purposive sampling, which may lead to over- or underrepresentation of certain expert groups. The two-stage selection strategy proposed by Selb et al. [4], which includes a random sampling of experts, would allow for control selection bias and obtain more balanced samples across regions.

Regarding the size of the expert group, there is no consensus on a minimum or maximum number of participants, which may vary depending on the study's objective [117]. The results of a bootstrap simulation study suggest that response stability over multiple rounds can be achieved even with 23 participants in the case of homogeneous expert groups [118]. The studies analyzed in our review had an average sample size of 77 experts, suggesting that researchers conducting ICF-CS validation studies should give greater consideration to the issue of response stability.

The reliability of the instruments used (item 9 in the AXIS tool) in Delphi studies is a significant concern. Only two studies explicitly reported having conducted a pilot of the questionnaire [36, 63], while a few others reported using questions from a previously published study. Some researchers also acknowledged the potential impact that altering the original questions or administration method might have on the study results [43, 63]. It is imperative to provide a comprehensive explanation of how a Delphi questionnaire was constructed and piloted to ensure the reliability and validity of study findings. Moreover, if published versions of a questionnaire were used, this should be made explicit in the methodology section [119, 120, 121]. By transparently addressing these methodological aspects, researchers can enhance the robustness of their Delphi studies.

Two crucial aspects regarding measurement accuracy and reliability merit attention. First, some authors argue that a robust methodological justification for consensus points among experts is essential, as these decisions often rely on practical criteria and previous research [122]. Reviews of Delphi studies have revealed substantial variability and inadequate reporting in defining consensus [123, 124]. Such variability can

introduce ambiguity and hinder the comparison of results. Most reviewed studies consistently used a 75% agreement threshold for consensus. However, the absence of accepted guidelines for ICF-CS consensus thresholds remains challenging. This ongoing ambiguity undermines the reliability and interpretability of results in Delphi studies. Consequently, further research and discussions may be needed to establish precise conventions in this vital aspect of Delphi methodology.

The second aspect to consider is that some studies have reported suboptimal levels of agreement among coders, with a few neglecting to incorporate statistical measures for assessing inter-rater agreement during the linking procedure. These omissions carry the potential to introduce bias into the results. Similar to our observations regarding interview and focus group studies, several authors have articulated concerns about the potential influence of coders' professional backgrounds or native languages on the Delphi study linking process. Indeed, the inherent subjectivity of the linking process, the possibility of divergent linking decisions, and the impact of diverse professional backgrounds all contribute to the complexity of the procedure, requiring further exploration and standardization efforts. Notably, researchers in the studies reviewed here demonstrated transparency in acknowledging the limitations and uncertainties associated with their linking processes. This transparency serves to enhance the interpretability and applicability of their findings.

One strategy to increase inter-coder agreement is to include a Delphi pilot in which thematic coding and linking tests can be conducted until an optimal level of inter-coder reliability is reached. For authors who choose to calculate kappa statistics, it is suggested that the percentage agreement should also be used, as the latter is a more accurate predictor of reliability [125]. Using both indices will increase the precision of the calculated agreement reliability in the linking process.

Finally, a notable finding from application of the AXIS tool was that half of the Delphi studies failed to mention obtaining informed consent or ethical committee approval. Delphi studies frequently omit this information, possibly because researchers consider they are conducting an opinion survey rather than a clinical study. However, Delphi studies are subject to the same ethical considerations as survey research [117], which include reporting research details and maintaining quasi-anonymous participation, among others. Thus, requesting informed consent from participants should be a requirement.

Strengths and limitations

The present review has certain limitations that should be considered. First, it is confined to literature published in English, potentially omitting valuable contributions in other languages and leading to biased results. Second, the fact that five of the studies reviewed only validated specific ICF-CS components (rather than the complete ICF-CS) may affect our conclusions concerning interview and ICF-CS questionnaire studies. These component-focused studies, driven by clinical interests in treatment goal indicators, may not fully represent the broader ICF-CS context. Consequently, their findings and recommendations may not be directly applicable to studies validating the complete ICF-CS or assessing a more comprehensive range of functioning and disability categories. These differences in focus and objectives must be taken into account when interpreting our findings.

Despite these limitations, the main and great strength of this review is that it analyzes the methodological quality not merely of ICF-CS validation studies but specifically those employing a cross-sectional research design, a combination that has not previously been studied. In addition, by applying a tool to assess the quality and risk of bias in cross-sectional studies, we were able to simultaneously analyze the use of qualitative, quantitative, and mixed methods to validate ICF-CSs, which aligns with an integrative view of validity evidence. Furthermore, the search criteria, the broad representation of databases in the health area, and the use of different methods to analyze methodological limitations allow us to establish solid conclusions regarding the practices used in carrying out ICF-CS validation studies with cross-sectional designs.

Conclusions

Our results indicate that research relating to ICF-CSs is lagging behind in the WHO regions of South-East Asia, the Eastern Mediterranean and, above all, Africa. Given that these regions experience more pronounced economic and developmental difficulties than do Europe, the Americas, and the Western Pacific regions, it is vital that they have access to validated instruments for assessing people's functioning, this being a crucial step in responding to their specific needs. Research efforts in these regions therefore need to be actively encouraged.

Comparison of our findings with those obtained a decade ago by Wiegand et al. [8] shows that ICF-CS studies limited to a particular context (i.e., single-center studies)

persist. While cross-sectional validation ICF-CS studies have progressed by incorporating multicenter approaches within a single country, there is a pressing need for more extensive multicenter studies involving several WHO regions. International multicenter research of this kind will enable a more precise analysis of the influence of cultural factors on different ICF-CS components and categories.

Our systematic review found that studies generally provided evidence of content validity and validity based on relationships with other variables. However, limited attention has been paid to assessing differences between relevant groups/subgroups, often due to small sample sizes [104, 108]. Previous research has highlighted the potential impact of gender, age, ethnicity, and varying health severity on functioning in different health conditions [126, 127, 128, 129, 130]. Therefore, these variables should be more comprehensively considered in the validation process [108], provided that samples exhibit sufficient variability.

Recommendations and future research

To ensure a globally representative ICF-CS validation process, we urgently advocate for the diversification of samples and research centers, particularly by actively involving underrepresented WHO regions. Encouraging international multicenter research efforts, encompassing participants from diverse WHO regions, is paramount to gaining a more nuanced understanding of how cultural factors influence ICF-CS categories. The inclusion of caregivers in future research is also essential to gain a different point of view regarding the functioning of people with specific health conditions.

Future cross-sectional validation studies should aim to enhance the characterization of non-responders and report response rates more comprehensively. Understanding the characteristics of non-respondents or refusals is crucial, as they may introduce bias into the results. To address this, we recommend the adoption of established guidelines for managing non-responders, drawing insights from sources such as Bose [131] for questionnaire studies, the COREQ guide [111] for qualitative studies, and Keeney [117] for Delphi studies.

It is also imperative to establish clear data supported cut-off criteria for category validation. In this regard, we recommend including sensitivity analyses to determine optimal cut-off values, and in their absence all categories should be analyzed. Robust sample size justification, particularly in questionnaire studies, and the standardization of

category scoring are likewise crucial steps toward enhancing result validity and reliability.

In qualitative studies employing interviews and focus groups, meticulous attention to sample size justification, factoring in group stratification and the underlying theoretical framework, is recommended. Researchers are also urged to enhance transparency and document their chosen perspective for saturation and linking. Ideally, a pilot test of the questions, coding, and linking procedures would be conducted [111]. This would help to determine the relevance and clarity of questions and identify potential environmental characteristics that may affect information collection, such as instruction clarity and the training of group leaders. A valuable guide to reporting interview and focus groups studies is the Consolidated Criteria for Reporting Qualitative Research (COREQ) [111]. However, interviewers or moderators should also receive training in interviewing and focus group techniques and possess specific knowledge of the ICF-CS.

We also recommend incorporating a pilot phase for both the questionnaire and inter-coder agreement assessments to enhance the quality of Delphi validation studies. Implementing a two-stage selection strategy involving random expert sampling could mitigate selection bias and yield more balanced regional samples. Expanding the application to languages other than English should also be considered, and informed consent from participants should be obtained. The CREDES guide [121] for conducting Delphi studies can be valuable for improving research reporting.

Methodologically, it is recommended that authors make explicit the type of sampling used in validation studies. We found in this review that this was not always the case, and we often had to infer the sampling technique from the information presented in the article, such as the description of the sample or the data collection process, although neither was this information always entirely clear or sufficient. This issue has a number of methodological implications. One is that it impacts the generalizability and replicability of data. In addition, when the sampling method must be inferred from other information, there is a risk that the specific approach used will be mistaken for a related but more general method. For example, it may be inferred from the information provided that the authors of a study used purposive sampling, when it fact they used a more specific approach that falls within this broad category, such as maximum variation sampling, stratified purposive sampling or criterion sampling [11].

Considering the methodological limitations identified within the scope of this review, we believe there is a need for replication studies to be conducted in various contexts.

We also endorse the implementation of qualitative and quantitative research syntheses, as these methodologies effectively amalgamate data derived from cross-sectional studies, thus enhancing the comprehensiveness and reliability of evidence. It is likewise of paramount importance to encourage the development of multicenter projects characterized by well-balanced sample sizes and dedicated response process validity studies. Lastly, there is a need for systematic reviews that analyze the methodological characteristics and reporting quality of ICF-CS validation studies employing other research designs, including longitudinal research, case-control and cohort studies, and secondary data analyses. This would provide furthermore detailed insights into the validity of ICF-CSs.

Acknowledgments

This study was supported by Spain's Ministry of Economy and Competitiveness (Grant PSI2015–67984-R) and by the Agency for the Management of University and Research Grants of the Government of Catalonia (grants 2021SGR01071 and 2017FI_B_00817). Erika Arias and Manuel Rojas would like to sincerely thank Alexander Calderón, statistician from the Universidad Nacional de Colombia, for answering our inquiries about the topic of sampling types and Sandra Camargo from Purdue University, for her response to our questions about the Delphi methodology, which enriched this article.

Data availability statement

The data presented in this study are available on request from the corresponding author's email address.

Declaration of interest

The authors report no conflicts of interest.

Disclosure statement

The authors declare that they have no competing interests.

References

[1] Stucki G, Cieza A, Melvin J. The International Classification of Functioning, Disability and Health (ICF): A unifying model for the conceptual description of the rehabilitation strategy. J Rehabil Med. 2007;39(4):279–285. doi: 10.2340/16501977-0041

[2] Selb M, Cieza A. ICF Core Sets. In: Bickenbach J, Cieza A, Selb M, Stucki G, eds. *ICF core sets: Manual for clinical practice*. 2nd ed. Hogrefe Publishing; 2021 p. 9–21.

[3] ICF Research Branch – ICF core sets. Icf-research-branch.org. [cited 2023 Mar 29]. Available from: http://www.icf-research-branch.org/icf-core-sets.

[4] Selb M, Escorpizo R, Kostanjsek N, et al. A guide on how to develop an International Classification of Functioning, Disability and Health Core Set. Eur J Phys Rehabil Med. 2015;51(1):105–117.

[5] Chan EKH. Standards and guidelines for validation practices: Development and evaluation of measurement instruments. In: Zumbo BD, Chan EKH, eds. *Validity and validation in social, behavioral, and health sciences*. Springer International Publishing; 2014. p. 9–24.

[6] Karlsson E, Gustafsson J. Validation of the International Classification of Functioning, Disability and Health (ICF) core sets from 2001 to 2019: A scoping review. Disabil Rehabil. 2022;44(14):3736–3748. doi: 10.1080/09638288.2021.1878562

[7] MacDermid JC. ICF linking and cognitive interviewing are complementary methods for optimizing content validity of outcome measures: An integrated methods review. Front Rehabil Sci. 2021;2:702596. doi: 10.3389/fresc.2021.702596

[8] Wiegand NM, Belting J, Fekete C, et al. All talk, no action? The global diffusion and clinical implementation of the International Classification of Functioning, Disability, and Health. Am J Phys Med Rehabil. 2012;91(7):550–560. doi: 10.1097/PHM.0b013e31825597e5

[9] Grill E, Stucki G. Criteria for validating comprehensive ICF Core Sets and developing brief ICF Core Set versions. J Rehabil Med. 2011;43(2):87–91. doi: 10.2340/16501977-0616

[10] Lavrakas PJ. *Encyclopedia of survey research methods*. Sage Publications; 2008.

[11] Bryman A. *Social research methods*. 5th ed. Oxford University Press; 2016.

[12] Hackshaw A. *A concise guide to observational studies in healthcare*. Wiley-Blackwell; 2015.

[13] Page MJ, Moher D, Bossuyt PM, et al. PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. BMJ. 2021;372:n160. doi: 10.1136/bmj.n160

[14] Downes MJ, Brennan ML, Williams HC, et al. Development of a critical appraisal tool to assess the quality of cross-sectional studies (AXIS). BMJ Open. 2016;6(12): e011458. doi: 10.1136/bmjopen-2016-011458.

[15] Kvale S. *Interviews: An introduction to qualitative research interviewing*. Sage Publications, Inc; 1996.

[16] Coenen M, Cieza A, Stamm TA, et al. Validation of the International Classification of Functioning, Disability and Health (ICF) Core Set for rheumatoid arthritis from the patient perspective using focus groups. Arthritis Res Ther. 2006;8(4):R84. doi: 10.1186/ar1956

[17] Aiachini B, Cremascoli S, Escorpizo R, et al. Validation of the ICF Core Set for Vocational Rehabilitation from the perspective of patients with spinal cord injury using focus groups. Disabil Rehabil. 2016;38(4):337–345. doi: 10.3109/09638288.2015.1041611

[18] Anner J, Brage S, Donceel P, et al. Validation of the EUMASS Core Set for medical evaluation of work disability. Disabil Rehabil. 2013;35(25):2147–2156. doi: 10.3109/09638288.2013.771709

[19] Bagraith KS, Strong J, Meredith PJ, et al. What do clinicians consider when assessing chronic low back pain? A content analysis of multidisciplinary pain centre team assessments of functioning, disability, and health. Pain. 2018;159(10):2128–2136. doi: 10.1097/j.pain.0000000000001285

[20] Cooney M, Galvin R, Connolly E, et al. The International Classification of Functioning (ICF) core set for breast cancer from the perspective of women with the condition. Disabil Rehabil. 2013;35(9):740–748. doi: 10.3109/09638288.2012.707742

[21] Glaessel A, Kirchberger I, Stucki G, et al. Does the Comprehensive International Classification of Functioning, Disability and Health (ICF) Core Set for Breast Cancer capture the problems in functioning treated by physiotherapists in women with breast cancer? Physiotherapy. 2011;97(1):33–46. doi: 10.1016/j.physio.2010.08.010

[22] Raji P, Hassani Mehraban A, Aliabadi F, et al. Content validity of the comprehensive ICF Core Set for children with cerebral palsy aged 0-6 years: Iranian occupational therapists perspective. Iran J Child Neurol. 2018;12(3):40–58.

[23] Liu B, Wang P, Zhou M, et al. Brief ICF core set for obstructive pulmonary diseases: Validation from the perspective of Chinese respiratory physicians. J Shanghai Jiatong Univ. 2021;26(1):25–32. doi: 10.1007/s12204-021-2257-9

[24] Zhang Y, Meng X, Shen Y, et al. The reliability and validity of the Brief ICF core set in patients with chronic obstructive pulmonary disease. Int J Chron Obstruct Pulmon Dis. 2021;16:2077–2087. doi: 10.2147/COPD.S306410

[25] Guo C, Liu Y, Hao S, et al. The reliability and validity of the “activity and participation” component in the Brief ICF Core Set for chronic obstructive pulmonary diseases based on Rasch analysis. Int J Chron Obstruct Pulmon Dis. 2020;15:1191–1198. doi: 10.2147/COPD.S249704

[26] Lage SM, Jácome C, Oliveira A, et al. Validation of the International Classification of Functioning, Disability and Health Core Set for obstructive pulmonary diseases in the perspective of adults with asthma. Disabil Rehabil. 2020;42(1):86–92. doi: 10.1080/09638288.2018.149315944

[27] Huang J, Reinhardt JD, Dai R, et al. Validation of the Brief International Classification of Functioning, Disability, and Health Core Set for obstructive pulmonary disease in the Chinese context. Chron Respir Dis. 2019;16:1-14. doi: 10.1177/1479973119843648

[28] Züge CH, Oliveira MR, Silva ALG da, et al. Understanding the functionality of people concerned by chronic obstructive pulmonary disease (COPD) under the perspective and validation of the Comprehensive ICF Core Set of the International Classification of Functionality. Brazilian Journal of Occupational Therapy. 2019;27(1):27–34. doi: 10.4322/2526-8910.ctoao1582

[29] Marques A, Jácome C, Gonçalves A, et al. Validation of the Comprehensive ICF Core Set for obstructive pulmonary diseases from the patient’s perspective. Int J

Rehabil Res. 2014;37(2):152–158. doi: 10.1097/MRR.0000000000000049

[30] Jobst A, Kirchberger I, Cieza A, et al. Content validity of the Comprehensive ICF Core Set for chronic obstructive pulmonary diseases: An international Delphi survey. Open Respir Med J. 2013;7(1):33–45. doi: 10.2174/1874306401307010033

[31] Marques A, Jácome C, Gabriel R, et al. Comprehensive ICF core set for obstructive pulmonary diseases: Validation of the activities and participation component through the patient's perspective. Disabil Rehabil. 2013;35(20):1686–1691. doi: 10.3109/09638288.2012.750691

[32] Rauch A, Kirchberger I, Stucki G, et al. Validation of the Comprehensive ICF Core Set for obstructive pulmonary diseases from the perspective of physiotherapists. Physiother Res Int. 2009;14(4):242–259. doi: 10.1002/pri.448

[33] Kurtaiş Y, Őztuna D, Genç A, et al. Reliability, construct validity and measurement potential of the International Classification of Functioning, Disability and Health Comprehensive Core Set for chronic widespread Pain, Journal of Musculoskeletal Pain. 2013;21(3):231-243. doi: 10.3109/10582452.2013.827772

[34] Hieblinger R, Coenen M, Stucki G, et al. Identification of essential elements of functioning in chronic widespread pain based on a statistical approach. Am J Phys Med Rehabil. 2011;90(12):979–991. doi: 10.1097/PHM.0b013e318228c89c

[35] Hieblinger R, Coenen M, Stucki G, et al. Validation of the International Classification of Functioning, Disability and Health Core Set for chronic widespread pain from the perspective of fibromyalgia patients. Arthritis Res Ther. 2009;11(3):R67. doi: 10.1186/ar2696

[36] Wildeboer AT, Stallinga HA, Roodbol PF. Validation of the International Classification of Functioning, Disability and Health (ICF) core set for diabetes mellitus from nurses' perspective using the Delphi method. Disabil Rehabil. 2022;44(2):210–218. doi: 10.1080/09638288.2020.1763485

[37] Tsutsui H, Ojima T, Ozaki N, et al. Validation of the Comprehensive International Classification of Functioning, Disability and Health (ICF) Core Set for diabetes mellitus in patients with diabetic nephropathy. Clin Exp Nephrol. 2015;19(2):254–263. doi: 10.1007/s10157-014-0983-0

[38] Awad H, Alghadir A. Validation of the Comprehensive International Classification of Functioning, Disability and Health Core Set for diabetes mellitus: physical therapists' perspectives. Am J Phys Med Rehabil. 2013;92(11):968–979. doi: 10.1097/PHM.0b013e31829b4a6d

[39] Abdullah MF, Nor NM, Mohd Ali SZ, et al. Validation of the Comprehensive ICF Core Sets for diabetes mellitus: A Malaysian perspective. Ann Acad Med Singapore. 2011;40(4):168–178. doi: 10.47102/annals-acadmedsg.v40n4p168

[40] Kirchberger I, Coenen M, Hierl FX, et al. Validation of the International Classification of Functioning, Disability and Health (ICF) Core Set for diabetes mellitus from the patient perspective using focus groups. Diabet Med. 2009;26(7):700–707. doi: 10.1111/j.1464-5491.2009.02762.x

[41] Kus S, Oberhauser C, Cieza A. Validation of the brief International Classification of Functioning, Disability, and Health (ICF) Core Set for hand conditions. J Hand Ther. 2012;25(3):274–287. doi: 10.1016/j.jht.2012.02.003

[42] Kus S, Dereskewitz C, Wickert M, et al. Validation of the Comprehensive International Classification of functioning, Disability and Health (ICF) Core Set for hand conditions. Hand Ther. 2011;16(3):58–66. doi: 10.1258/ht.2011.011011

[43] Engelbarts M, Schuster V, Kisser U, et al. The Comprehensive ICF Core Set for head and neck cancer: A Delphi consensus survey among German speaking speech and language therapists. Eur Arch Otorhinolaryngol. 2017;274(6):2589–2599. doi: 10.1007/s00405-017-4494-y

[44] Leib A, Cieza A, Tschiesner U. Perspective of physicians within a multidisciplinary team: Content validation of the Comprehensive ICF Core Set for head and neck cancer. Head Neck. 2012;34(7):956–966. doi: 10.1002/hed.21844

[45] Becker S, Kirchberger I, Cieza A, et al. Content validation of the Comprehensive ICF Core Set for head and neck cancer (HNC): The perspective of psychologists. Psycho-oncology. 2010;19(6):594–605. doi: 10.1002/pon.1608

[46] Tschiesner U, Linseisen E, Becker S, et al. Content validation of the International Classification of Functioning, Disability and Health Core Sets for head and neck cancer: A multicentre study. J Otolaryngol Head Neck Surg. 2010;39(6):674–687

[47] Karlsson E, Mäki-Torkko E, Widén S, et al. Validation of the Brief International Classification of Functioning, Disability and Health (ICF) Core Set for hearing loss: An international multicentre study. Int J Audiol. 2021;60(6):412–420. doi: 10.1080/14992027.2020.1846088

[48] Riberto M, Chiappetta LM, Lopes KA, et al. A transversal multicenter study assessing functioning, disability and environmental factors with the comprehensive ICF core set for low back pain in Brazil. Eur J Phys Rehabil Med. 2014;50(2):153–160.

[49] Glocker C, Kirchberger I, Gläßel A, et al. Content validity of the Comprehensive International Classification of Functioning, Disability and Health (ICF) Core Set for low back pain from the perspective of physicians: A Delphi survey. Chronic Illn. 2013;9(1):57–72. doi: 10.1177/1742395312451280

[50] Røe C, Bautz-Holter E, Cieza A. Low back pain in 17 countries. A Rasch analysis of the ICF Core Set for low back pain. Int J Rehabil Res. 2013;36(1):38–47. doi: 10.1097/MRR.0b013e328357883c

[51] Kirschneck M, Kirchberger I, Amann E, et al. Validation of the Comprehensive ICF Core Set for low back pain: The perspective of physical therapists. Man Ther. 2011;16(4):364–372. doi: 10.1016/j.math.2010.12.011

[52] Öztuna D, Yanik B, Kutlay Ş, et al. Psychometric properties of the ICF core set for low back pain and its clinical use. Turk J Rheumatol. 2011;26(1):44–52. doi: 10.5606/tjr.2011.007

[53] Røe C, Sveen U, Geyh S, et al. Construct dimensionality and properties of the categories in the ICF Core Set for low back pain. J Rehabil Med. 2009;41(6):429–437. doi: 10.2340/16501977-0368

[54] Røe C, Sveen U, Cieza A, et al. Validation of the Brief ICF Core Set for low back pain from the Norwegian perspective. Eur J Phys Rehabil Med. 2009;45(3):403–414.

[55] Bautz-Holter E, Sveen U, Cieza A, et al. Does the International Classification of Functioning, Disability and Health (ICF) Core Set for low back pain cover the patients' problems? A cross-sectional content-validity study with a Norwegian population. Eur J Phys Rehabil Med. 2008;44(4):387–397

[56] Ortabozkoyun Ö, Kesikburun B, Köseoğlu BF, et al. Validation of the Comprehensive ICF Core Set for multiple sclerosis: A Turkish clinical perspective. Turk J Phys Med Rehab 2023;70(1):17-29. doi: 10.5606/tftrd.2023.11907

[57] Renom M, Conrad A, Bascuñana H, et al. Content validity of the Comprehensive ICF Core Set for multiple sclerosis from the perspective of speech and language therapists. Int J Lang Commun Disord. 2014;49(6):672–686. doi: 10.1111/1460-6984.12086

[58] Karhula ME, Kanelisto KJ, Ruutiainen J, et al. The activities and participation categories of the ICF Core Sets for multiple sclerosis from the patient perspective. Disabil Rehabil. 2013;35(6):492–497. doi: 10.3109/09638288.2012.702845

[59] Berno S, Coenen M, Leib A, et al. Validation of the Comprehensive International Classification of Functioning, Disability, and Health Core Set for multiple sclerosis from the perspective of physicians. J Neurol. 2012;259(8):1713–1726. doi: 10.1007/s00415-011-6399-9

[60] Conrad A, Coenen M, Schmalz H, et al. Validation of the Comprehensive ICF Core Set for multiple sclerosis from the perspective of physical therapists. Phys Ther. 2012;92(6):799–820. doi: 10.2522/ptj.20110056

[61] Conrad A, Coenen M, Schmalz H, et al. Validation of the Comprehensive ICF Core Set for multiple sclerosis from the perspective of occupational therapists. Scand J Occup Ther. 2012;19(6):468–487. doi: 10.3109/11038128.2012.665475

[62] Hernández-Lázaro H, Mingo-Gómez MT, Jiménez-Del-Barrio S, et al. Validation of the International Classification of Functioning, Disability, and Health (ICF) Core Set for post-acute musculoskeletal conditions in a primary care physiotherapy setting from the perspective of patients using focus groups. Disabil Rehabil. 2023;5:1–8. doi: 10.1080/09638288.2023.2251392

[63] Hernández-Lazaro H, Mingo-Gómez MT, Ceballos-Laita L, et al. Validation of the International Classification of Functioning, Disability, and Health (ICF) Core Sets for musculoskeletal conditions in a primary health care setting from physiotherapists' perspective using the Delphi method. Disabil Rehabil. 2022;45(15):2458-2468. doi: 10.1080/09638288.2022.2096128

[64] Weigl M, Wild H. European validation of the Comprehensive International Classification of Functioning, Disability and Health Core Set for osteoarthritis from the perspective of patients with osteoarthritis of the knee or hip. Disabil Rehabil. 2018;40(26):3104–3112. doi: 10.1080/09638288.2017.1377295

[65] Oberhauser C, Escorpizo R, Boonen A, et al. Statistical validation of the Brief International Classification of Functioning, Disability and Health Core Set for osteoarthritis based on a large international sample of patients with osteoarthritis. Arthritis Care Res. 2013;65(2):177–186. doi: 10.1002/acr.21775

[66] Bossmann T, Kirchberger I, Glaessel A, et al. Validation of the Comprehensive ICF Core Set for osteoarthritis: The perspective of physical therapists. Physiotherapy. 2011;97(1):3–16. doi: 10.1016/j.physio.2009.11.011

[67] Kurtaiş Y, Oztuna D, Küçükdeveci AA, et al. Reliability, construct validity and measurement potential of the ICF Comprehensive Core Set for osteoarthritis. BMC Musculoskelet Disord. 2011;12:255. doi: 10.1186/1471-2474-12-255

[68] Xie F, Lo NN, Lee HP, et al. Validation of the International Classification of Functioning, Disability, and health (ICF) Brief Core Set for osteoarthritis. Scand J Rheumatol. 2008;37(6):450–461. doi: 10.1080/03009740802116216

[69] Xie F, Lo NN, Lee HP, et al. Validation of the Comprehensive ICF Core Set for osteoarthritis (OA) in patients with knee OA: A Singaporean perspective. J Rheumatol. 2007;34(11):2301–2307.

[70] Xie F, Thumboo J, Fong KY, et al. Are they relevant? A critical evaluation of the International Classification of Functioning, Disability, and Health Core Sets for

osteoarthritis from the perspective of patients with knee osteoarthritis in Singapore. Ann Rheum Dis. 2006;65(8):1067–1073. doi: 10.1136/ard.2005.043067

[71] Koehler B, Kirchberger I, Glaessel A, et al. Validation of the International Classification of Functioning, Disability and Health Comprehensive Core Set for osteoporosis: The perspective of physical therapists. J Geriatr Phys Ther. 2011;34(3):117–130. doi: 10.1519/JPT.0b013e31820aa990

[72] Monteiro IO, Oliveira NSC, Ruaro JA, et al. Convergent validity and reproducibility of the International Classification of Functioning, Disability and Health (ICF) Core Set for the physical health of community-dwelling older adults. Braz J Phys Ther. 2021;25:563–572. doi: 10.1016/j.bjpt.2021.02.007

[73] Oliveira NSC, Monteiro IO, Ruaro JA, et al. Assessment of the concurrent validity of the ICF Core Set to classify the physical health of community-dwelling older adults in relation to self-rated health. J Geriatr Phys Ther. 2023;46(1):71–81. doi: 10.1519/JPT.0000000000000314

[74] Gebhardt C, Kirchberger I, Stucki G, et al. Validation of the Comprehensive ICF Core Set for rheumatoid arthritis: The perspective of physicians. J Rehabil Med. 2010;42(8):780–788. doi: 10.2340/16501977-0599

[75] Kirchberger I, Cieza A, Stucki G. Validation of the Comprehensive ICF Core Set for rheumatoid arthritis: The perspective of psychologists. Psychol Health. 2008;23(6):639–659. doi: 10.1080/14768320701244076

[76] Kirchberger I, Stamm T, Cieza A, et al. Does the Comprehensive ICF Core Set for rheumatoid arthritis capture occupational therapy practice? A content-validity study. Can J Occup Ther. 2007;74(5_suppl):267–280. doi: 10.1177/000841740707405s05

[77] Kirchberger I, Glaessel A, Stucki G, et al. Validation of the Comprehensive International Classification of Functioning, Disability and Health Core Set for rheumatoid arthritis: The perspective of physical therapists. Phys Ther. 2007;87(4):368–384. doi: 10.2522/ptj.20050237

[78] Stamm TA, Cieza A, Coenen M, et al. Validating the International Classification of Functioning, Disability and Health Comprehensive Core Set for rheumatoid arthritis from the patient perspective: A qualitative study. Arthritis Rheum. 2005;53(3):431–439. doi: 10.1002/art.21159

[79] Nuño L, Guilera G, Rojo E, et al. An integrated account of expert perspectives on functioning in schizophrenia. J Clin Med. 2021;10(18):4223. doi: 10.3390/jcm10184223

[80] Nuño L, Guilera G, Bell M, et al. Occupational therapists' perspective on the International Classification of Functioning, Disability and Health Core Sets (ICF-CS) for schizophrenia. Am J Occup Ther. 2021;75(2):7502205060p1-7502205060p10. doi: 10.5014/ajot.2021.041509

[81] Nuño L, Guilera G, Solomon P, et al. The perspective of social workers on functioning for individuals with schizophrenia: A Delphi study. J Soc Social Work Res. 2021;12:591–617. doi: 10.1086/715893

[82] Nuño L, Barrios M, Vancampfort D, et al. Functioning in schizophrenia: A Delphi study covering the perspective of physiotherapists. Disabil Rehabil. 2021;43(26):3733–3740. doi: 10.1080/09638288.2020.1748729

[83] Nuño L, Guilera G, Coenen M, et al. Functioning in schizophrenia from the perspective of psychologists: A worldwide study. PLoS One. 2019;14(6):e0217936. doi: 10.1371/journal.pone.0217936

[84] Nuño L, Barrios M, Moller MD, et al. An international survey of psychiatric mental health nurses on the content validity of the International Classification of Functioning, Disability and Health Core Sets for schizophrenia. Int J Ment Health Nurs. 2019;28(4):867–878. doi: 10.1111/inm.12586

[85] Nuño L, Barrios M, Rojo E, et al. Validation of the ICF Core Sets for schizophrenia from the perspective of psychiatrists: An international Delphi study. J Psychiatr Res. 2018;103:134–141. doi: 10.1016/j.jpsychires.2018.05.012

[86] Herrmann KH, Kirchberger I, Stucki G, et al. The Comprehensive ICF Core Sets for spinal cord injury from the perspective of physical therapists: A worldwide validation study using the Delphi technique. Spinal Cord. 2011;49(4):502–514. doi: 10.1038/sc.2010.155

[87] Herrmann KH, Kirchberger I, Stucki G, et al. The Comprehensive ICF Core Sets for spinal cord injury from the perspective of occupational therapists: A worldwide validation study using the Delphi technique. Spinal Cord. 2011;49(5):600–613. doi: 10.1038/sc.2010.168

[88] Tatlı HU, Köseoğlu BF, Özcan DS, et al. Validation and application of the International Classification of Functioning Core Set for spinal cord injury in the Turkish patients. Turk J Phys Med Rehabil. 2019;65(3):244–258. doi: 10.5606/tftrd.2019.3045

[89] Glässel A, Coenen M, Kollerits B, et al. Content validation of the International Classification of Functioning, Disability and Health Core Set for stroke from gender perspective using a qualitative approach. Eur J Phys Rehabil Med. 2014;50(3):285–299.

[90] Paanalahti M, Alt Murphy M, Lundgren-Nilsson Å, et al. Validation of the Comprehensive ICF Core Set for stroke by exploring the patient's perspective on functioning in everyday life: A qualitative study. Int J Rehabil Res. 2014;37(4):302–310. doi: 10.1097/mrr.0000000000000070

[91] Wang P, Li H, Guo Y, et al. The feasibility and validity of the Comprehensive ICF Core Set for stroke in Chinese clinical settings. Clin Rehabil. 2014;28(2):159–171. doi: 10.1177/0269215513496659

[92] Köseoğlu B, Nebahat Öken O, Serap Kibar S. Validation of the Comprehensive ICF Core Set for stroke in Turkish stroke patients. Turk Geriatri Dergisi. 2013;16(1):8-19.

[93] Riberto M, Lopes KAT, Chiappetta LM, et al. The use of the Comprehensive International Classification of Functioning, Disability and Health Core Set for stroke for chronic outpatients in three Brazilian rehabilitation facilities. Disabil Rehabil. 2013;35(5):367–374. doi: 10.3109/09638288.2012.694573

[94] Glässel A, Coenen M, Kollerits B, et al. Validation of the Extended ICF Core Set for stroke from the patient perspective using focus groups. Disabil Rehabil. 2012;34(2):157–166. doi: 10.3109/09638288.2011.593680

[95] Algurén B, Bostan C, Christensson L, et al. A multidisciplinary cross-cultural measurement of functioning after stroke: Rasch analysis of the Brief ICF Core Set for stroke. Top Stroke Rehabil. 2011;18 Suppl1(sup1):573–586. doi: 10.1310/tsr18s01-573

[96] Glässel A, Kirchberger I, Kollerits B, et al. Content validity of the Extended ICF Core Set for stroke: An international Delphi survey of physical therapists. Phys Ther. 2011;91(8):1211–1222. doi: 10.2522/ptj.20100262

[97] Algurén B, Lundgren-Nilsson A, Sunnerhagen KS. Functioning of stroke survivors: A validation of the ICF Core Set for stroke in Sweden. Disabil Rehabil.

2010;32(7):551–559. doi: 10.3109/09638280903186335

[98] Glässel A, Kirchberger I, Linseisen E, et al. Content validation of the International Classification of Functioning, Disability and Health (ICF) Core Set for stroke: The perspective of occupational therapists. Can J Occup Ther. 2010;77(5):289–302. doi: 10.2182/cjot.2010.77.5.5

[99] Lemberg I, Kirchberger I, Stucki G, et al. The ICF Core Set for stroke from the perspective of physicians: A worldwide validation study using the Delphi technique. Eur J Phys Rehabil Med. 2010;46(3):377–388.

[100] Kinoshita S, Abo M, Miyamura K, et al. Validation of the "Activity and participation" component of ICF Core Sets for stroke patients in Japanese rehabilitation wards. J Rehabil Med. 2016;48(9):764-768. doi: 10.2340/16501977-2126

[101] Chung P, Khan F, Judson R. Validation of the International Classification of Functioning, Disability and Health Core Sets for traumatic brain injury from Australian community patient perspectives. J Rehabil Med. 2021;53(7):jrm00218. doi: 10.2340/16501977-2860

[102] Kaech Moll VM, Escorpizo R, Portmann Bergamaschi R, et al. Validation of the Comprehensive ICF Core Set for vocational rehabilitation from the perspective of physical therapists: International Delphi survey. Phys Ther. 2016;96(8):1262–1275. doi: 10.2522/ptj.20150365

[103] Saunders B, Sim J, Kingstone T, et al. Saturation in qualitative research: Exploring its conceptualization and operationalization. Qual Quant. 2018;52(4):1893–1907. doi: 10.1007/s11135-017-0574-8

[104] Kang H. Sample size determination and power analysis using the G*Power software. J Educ Eval Health Prof. 2021;18:17. doi: 10.3352/jeehp.2021.18.17

[105] Boateng GO, Neilands TB, Frongillo EA, et al. Best practices for developing and validating scales for health, social, and behavioral research: A primer. Front Public Health. 2018;11:6:149 doi: 10.3389/fpubh.2018.00149

[106] Prodinger B, Stucki G, Coenen M, et al. The measurement of functioning using the International Classification of Functioning, Disability and Health: Comparing qualifier ratings with existing health status instruments. Disabil Rehabil. 2019;41(5):541–548. doi: 10.1080/09638288.2017.1381186

[107] Tennant A, Penta M, Tesio L, et al. Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model: The PRO-ESOR project. Med Care. 2004;42(1 Suppl):I37-I48. doi: 10.1097/01.mlr.0000103529.63132.77

[108] Wright J, Moghaddam N, Dawson DL. Cognitive interviewing in patient-reported outcome measures: A systematic review of methodological processes. Qual Psychol. 2021;8(1):2–29. doi: 10.1037/qup0000145

[109] American Educational Research Association, American Psychological Association, National Council on Measurement in Education, editors. Standards for educational and psychological testing. Lanham, MD: American Educational Research Association; 2014.

[110] Vasileiou K, Barnett J, Thorpe S, et al. Characterising and justifying sample size sufficiency in interview-based studies: Systematic analysis of qualitative health research over a 15-year period. BMC Med Res Methodol. 2018;18(1):148. doi: 10.1186/s12874-018-0594-7

[111] Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): A 32-item checklist for interviews and focus groups. Int J

Qual Health Care. 2007;19(6):349–357. doi: 10.1093/intqhc/mzm042

[112] Hennink MM, Kaiser BN, Weber MB. What influences saturation? Estimating sample sizes in focus group research. Qual Health Res. 2019;29(10):1483–1496. doi:10.1177/1049732318821692

[113] Hennink MM, Kaiser BN. Sample sizes for saturation in qualitative research: A systematic review of empirical tests. Soc Sci Med. 2022;292:114523. doi: 10.1016/j.socscimed.2021.114523

[114] Coenen M, Stamm TA, Stucki G, et al. Individual interviews and focus groups in patients with rheumatoid arthritis: A comparison of two qualitative methods. Qual Life Res. 2012;21:359–370. doi: 10.1007/s11136-011-9943-2

[115] Fayed N, Cieza A, Bickenbach JE. Linking health and health-related information to the ICF: A systematic review of the literature from 2001 to 2008. Disabil Rehabil. 2011;33(21-22):1941–1951. doi: 10.3109/09638288.2011.553704

[116] Cieza A, Fayed N, Bickenbach J, et al. Refinements of the ICF linking rules to strengthen their potential for establishing comparability of health information. Disabil Rehabil. 2019;41:574–583. doi:10.3109/09638288.2016.1145258

[117] Keeney S, Hasson F, McKenna HP. A critical review of the Delphi technique as a research methodology for nursing. Int J Nurs Stud. 2001;38(2):195–200. doi: 10.1016/s0020-7489(00)00044-4

[118] Akins RB, Tolson H, Cole BR. Stability of response characteristics of a Delphi panel: Application of bootstrap data expansion. BMC Med Res Methodol. 2005;5:37. doi: 10.1186/1471-2288-5-37

[119] Clibbens N, Walters S, Baird W. Delphi research: Issues raised by a pilot study. Nurse Res. 2012;19(2):37–44. doi: 10.7748/nr2012.01.19.2.37.c8907

[120] Keeney S, McKenna H, Hasson F. *The Delphi technique in nursing and health research*. Wiley-Blackwell; 2011.

[121] Jünger S, Payne SA, Brine J, et al. Guidance on Conducting and REporting DElphi Studies (CREDES) in palliative care: Recommendations based on a methodological systematic review. Palliat Med. 2017;31(8):684–706. doi: 10.1177/0269216317690685

[122] Barrios M, Guilera G, Nuño L, et al. Consensus in the Delphi method: What makes a decision change? Technol Forecast Soc Change. 2021;163:120484. doi: 10.1016/j.techfore.2020.120484

[123] Niederberger M, Spranger J. Delphi technique in health sciences: A map. Front Public Health. 2020;8:457. doi: 10.3389/fpubh.2020.00457

[124] Diamond IR, Grant RC, Feldman BM, et al. Defining consensus: A systematic review recommends methodologic criteria for reporting of Delphi studies. J Clin Epidemiol. 2014;67(4):401–409. doi: 10.1016/j.jclinepi.2013.12.002

[125] Zhao X, Feng GC, Ao SH, et al. Interrater reliability estimators tested against true interrater reliabilities. BMC Med Res Methodol. 2022;22(1):232. doi: 10.1186/s12874-022-01707-5

[126] Rojas M, Barrios M, Gómez-Benito J, et al. Functioning problems in persons with schizophrenia in the Russian context. Int J Environ Res Public Health. 2021;18(19):10276. doi: 10.3390/ijerph181910276

[127] Radosavljevic K, Dragovic-Lukic G, Nikolic D, et al. Gender and musculoskeletal comorbidity impact on physical functioning in elderly after hip fracture: The role of rehabilitation. Healthcare (Basel). 2020;8(1):31. doi: 10.3390/healthcare8010031

[128] Prat G, Escandell MJ, Garcia-Franco M, et al. Gender differences in social functioning in people with schizophrenia in psychosocial rehabilitation services using one-dimensional and multidimensional instruments. Compr Psychiatry. 2018;87:120–122. doi: 10.1016/j.comppsych.2018.10.004

[129] Fehrmann E, Kotulla S, Fischer L, et al. The impact of age and gender on the ICF-based assessment of chronic low back pain. Disabil Rehabil. 2019;41(10):1190–1199. doi: 10.1080/09638288.2018.1424950

[130] Norrefalk J-R, Borg K. The Functional Barometer – An analysis of a self-assessment questionnaire with ICF-coding regarding functional/activity limitations and quality of life due to pain: Differences in age gender and origin of pain. Scand J Pain. 2017;17:16–21. doi: 10.1016/j.sjpain.2017.06.004

[131] Bose J. Non-response bias analyses at the National Center for Education Statistics. In: *Achieving data quality in a statistical agency: A methodological perspective*. Proceedings of Statistics Canada Symposium 2001. 2001. p. 9.