COSMIN: Development and evaluation of a checklist to assess the methodological quality of studies on measurement properties

Mokkink, L.B.

2010

document version
Publisher's PDF, also known as Version of record

Summary

COSMIN

COSMIN is an acronym for COnsensus-based Standards for the selection of health Measurement INstruments. The COSMIN initiative aims to improve the selection of health measurement instruments. The research described in this dissertation contributes to the COSMIN initiative, and its initial focus is on evaluative health-related patient-reported outcomes (HR-PROs). We defined *evaluative* as instruments which are applied to measure HR-PROs in a longitudinal study to assess change over time.

The selection of a health measurement instrument should depend – among other things – on the quality of the instruments. To evaluate the quality of a measurement instrument, its measurement properties (e.g. reliability, measurement error or content validity) need to be assessed. The studies in which the measurement properties are being investigated should be methodologically sound. Studies of low methodological quality may give biased results, and consequently cannot be trusted. Therefore, the methodological quality of studies on measurement properties should be evaluated to detect potential bias.

This dissertation is about the development and evaluation of the COSMIN checklist, which can be used to assess the methodological quality of studies on one or more measurement properties. It contains standards that were developed in the COSMIN Delphi study. The research questions of the Delphi study were:

1. Which measurements properties should be included in the assessment of evaluative HR-PROs, and how should they be defined?
2. How should these measurement properties be assessed in terms of study design and statistical analysis? (i.e. standards)
3. Which criteria should be applied to define what good measurement properties are?

A systematic review of systematic reviews of measurement properties was conducted and used as input for the Delphi study. Subsequently, the inter-rater reliability of the COSMIN checklist was assessed.

In Chapter 1 we gave an introduction about guidelines for quality assessment, and presented the research question.

In Chapter 2 we presented the design of the COSMIN Delphi study. The research questions (see above) and the design of the Delphi study were described. The COSMIN Delphi study contains a preparation phase, in which a systematic literature review was conducted to search for systematic reviews of evaluative health status measurement instruments. An inventory was made of definitions of measurement properties and existing methodological criteria lists. The chapter described how experts were selected. Experts participating in the Delphi study had a background in epidemiology, statistics, psychology and clinical medicine. Four written rounds were planned to reach consensus on the research questions, and a field-testing phase, to test the inter-rater reliability of the COSMIN checklist. In each Delphi

round the results of the previous round were presented in a feedback report. The panel was asked to rate their (dis)agreement about proposals on a 5-point scale. Consensus was considered to be reached when at least 67% of the panel agreed.

In Chapter 3 we undertook a systematic review of systematic reviews of measurement properties, with the aim to appraise the quality of the review process, to describe how authors assess the methodological quality of primary studies of measurement properties, and to describe how authors evaluate results of these studies. Literature searches were performed in PubMed, Embase, and PsycInfo. We included reviews which identified health status instruments used in an evaluative application and to report on the measurement properties of these instruments. Two independent reviewers selected the articles and extracted the data. We found 148 reviews on measurement properties of instruments measuring general health perceptions (43%), functional status (21%), symptoms (17%), biological and physiological processes (5%), or a combination of these concepts (14%). We concluded that during the last few years the number of such systematic reviews published has increased substantially. However, the methodological quality of these reviews leaves much to be desired. In 56% (n=83/148) of the reviews the methodological quality of the included studies was (partly) assessed by the authors of the reviews and (some of) the results were evaluated, i.e. standards and/or criteria of adequacy were applied to one or more measurement properties. Often a limited number of standards and/or criteria of adequacy were applied. In only seven reviews standards for each measurement property as well as criteria of adequacy for each measurement property were applied.

The Chapters 4-6 are based on the COSMIN Delphi study. The COSMIN Delphi study was an international Delphi-study with four written rounds. In this Delphi study we reached consensus on terminology, definitions, a taxonomy of measurement properties, and standards for evaluating measurement properties. In Appendix 4 the COSMIN taxonomy is presented, and in Appendix 5 the terminology and definitions can be found. A Figure of the 4-step procedure to complete the COSMIN checklist can be found in Appendix 6, and the COSMIN checklist can be found in Appendix 7.
We expect that this consensus will lead to a more uniform use of terms and definitions in the literature on measurement properties. Lack of consensus has lead to confusion about which measurement properties are relevant, which concepts are represented, and how these measurement properties should be assessed in terms of design requirements and preferred statistical methods. Terminology, definitions and standards used in the systematic reviews included in the systematic review described in Chapter 3 or found in the additional search in methodological literature, were used as input in the COSMIN Delphi study.

In Chapter 4 we described the consensus reached by the panel on terminology (percentage consensus ranging from 74% to 88%, except for one term (only 56% agreed on structural validity)), definitions (percentage consensus ranging from 68% to 88%), and a taxonomy of measurement properties.
Some results raised a lot of discussion. We described the discussions of the panel about the position of the measurement properties internal consistency and responsiveness in the taxonomy, the terminology for reliability and for structural validity, and the definitions of the measurement properties internal consistency, reliability, and responsiveness.
It was discussed whether or not internal consistency should be considered a separate measurement property, or an aspect of the measurement property reliability. It was decided that internal consistency was a separate measurement property in the domain reliability. Furthermore, it was decided that the domain responsiveness should be presented separately from the domain validity, and we discussed whether the domain responsiveness should consist of two measurement properties, i.e. construct responsiveness and criterion responsiveness, which are similar to construct validity and criterion validity. The panel disagreed with introducing new terms. Therefore, it is called responsiveness. The difference between standards for the situation in which a gold standard exists, and the situation when a gold standard is lacking, is now reflected in the standards.

We discussed whether we should use the term "reliability" or "reproducibility" for the term of the measurement property. Although the domain was also called "reliability", the panel preferred the term "reliability" for the measurement property. Next, we discussed whether we should use the term factorial validity or structural validity. Arguments against the term factorial validity were that it referred only to one of the methods to evaluate this aspect of construct validity, while structural validity referred to the purpose of this aspect. The panel did not reach consensus on the choice between the terms. Therefore, the Steering Committee decided to use the term structural validity.
To reach consensus on a definition for internal consistency, we started with a definition that tried to explain internal consistency and at the same time tried to reflect the difference between internal consistency and homogeneity. However, the panel preferred a definition that only explained internal consistency. The difference with homogeneity is now reflected in the standards.
The initially chosen definition of the measurement property reliability was not in agreement with the most preferred statistical methods, i.e. intraclass correlation coefficient (ICC) or Cohen's kappa. Therefore, another definition was proposed, and accepted by the panel.
The panel reached consensus on the definition of responsiveness, i.e. "the ability of an instrument to detect important change over time in the construct to be measured". However, it was decided to remove the word "important", because the importance of the detected

change is a separate issue that refers to the interpretation of the change score. In addition, the cut-off point between important change and non-important change is quite arbitrary.

In Chapter 5 we presented the COSMIN checklist, containing the standards for evaluating the methodological quality of studies on measurement properties on which we reached consensus in the Delphi study. To complete the COSMIN checklist a 4-step procedure should be followed: Step 1 is to determine which measurement properties are evaluated in an article; Step 2 is to determine whether Item Response Theory (IRT) is used in the article; Step 3 is to evaluate the methodological quality of the studies on the properties identified in step 1; Step 4 is to assess the generalisability of the results of the studies on the properties identified in step 1.
The COSMIN checklist contains twelve boxes. One box contains four items about general requirements for articles in which IRT methods are applied (i.e. step 2). Ten boxes can be used to assess whether a study meets the standards for good methodological quality (i.e. step 3). Nine of these boxes contain standards for measurement properties (internal consistency, reliability, measurement error, content validity, structural validity, hypotheses testing, cross-cultural validity, criterion validity, and responsiveness; ranging from 5-18 items), and one box contains standards for studies on interpretability (9 items). In addition, one box contains 8 items about general requirements for the generalisability of the results (i.e. step 4). For more details on the checklist, we refer to the COSMIN checklist manual (www.cosmin.nl).
The COSMIN checklist contains standards for evaluating the methodological quality of studies on measurement properties. Therefore the COSMIN checklist is not meant for evaluating the quality of the HR-PRO instrument itself. To assess the quality of the instrument, criteria for what constitutes good measurement properties should be applied to the results of a study on measurement properties. These criteria of adequacy are necessary to select the best measurement instrument. These should be developed in the future.

In Chapter 6 we described in more detail several items of the COSMIN checklist. We explained our choices for the included design requirements and preferred statistical methods against the background of existing literature. Herewith, we aim to contribute to a better understanding of the rationale behind the items, thereby enhancing the acceptance and use of the COSMIN checklist.
We explained that internal consistency is only relevant for measurement instruments of constructs that are based on reflective models, and that unidimensionality is a prerequisite for internal consistency.
Content validity should be assessed by making a judgment about the relevance and the comprehensiveness of the items.

Hypotheses-testing is one of the three aspects of construct validity. It concerns the relationships to scores of other instruments, or differences between relevant groups. Hypotheses-testing is an ongoing, iterative process, in which specific hypotheses should include an indication of the expected direction and magnitude of correlations or differences.
Criterion validity is defined as the degree to which the scores of a HR-PRO instrument are an adequate reflection of a "gold standard". The Delphi panel reached consensus that no gold standards exist for HR-PRO instruments, and decided that the only exception of a gold standard is when a shortened instrument is compared to the original long version.
Responsiveness is considered as a separate measurement property, but the panel agreed that the only difference between cross-sectional (construct and criterion) validity and responsiveness is that validity refers to the validity of a single score, and responsiveness refers to the validity of a change score. Because no gold standards exist for HR-PROs, responsiveness is appropriately measured by testing pre-specified hypotheses about the relations of changes in the questionnaires with changes in other measures. In Chapter 6 we also explained why parameters such as effect sizes (mean change score/SD baseline), and related measures, such as standardised response mean (mean change score/SD change score), and also paired t-test, Guyatt's responsiveness ratio (MIC/SD change score of stable patients), and minimal important change (MIC) are inappropriate measures for responsiveness.

The aim of the study described in Chapter 7 was to investigate the inter-rater reliability of each item of the COSMIN checklist. Therefore, we randomly selected 75 articles on measurement properties on HR-PROs. For each article the workload for assessing the article was determined, based on the number of measurement properties assessed, the number of instruments that were studied, the number of pages, and whether IRT was used. Eighty-eight participants each assessed the methodological quality of three articles, using the COSMIN checklist. Each participant received a low-workload article, a moderate-workload article, and a high-workload article. Kappa's (using one-way design) and percentage agreement were calculated for the items of the COSMIN checklist. Fifty-nine percent of the items showed poor kappa's (below 0.40), and 6% showed excellent kappa's (above 0.75). However, the percentage agreement was appropriate (i.e. above 80%) for 64% of the items. Reasons for low agreements were (1) a lack of dispersal of the scores of items, (2) a need for subjective judgement for several items, (3) poor reporting resulting in a lack of information needed to answer items, (4) inappropriate reading of the manual, or disagreement on items of the COSMIN checklist, (5) use of terminology or taxonomy by authors of the original studies that deviated from COSMIN, and (6) the difficult distinction between the standards for studies (i.e. referring to the methodological quality of the study) and criteria of adequacy (referring to the quality of the instrument). This applied especially for content validity.

To improve the agreement, we improved the manual. Furthermore, we recommend getting some experience in completing the COSMIN checklist before using it, e.g. in a systematic review. We strongly recommend using the taxonomy and terminology of the COSMIN checklist. For example, if authors compare their PRO to a commonly used PRO such as the SF-36, and they refer to this as criterion validity, we recommend considering this an evaluation of hypotheses-testing, and complete box F. When using the checklist in a systematic review of measurement instruments, we recommend that two raters first complete the checklist independently, and then reach consensus on one final rating. We advise the review team to agree beforehand on decisions to be made for the items that need a subjective judgement, and how to deal with lack of reporting in the original article. The inter-rater reliability of the consensus obtained by couples of reviewers might be higher, but this needs to be examined. Apart from the reliability of the COSMIN checklist, we recommend evaluating content validity, construct validity by hypotheses-testing, and interpretability of the COSMIN checklist, and we argue that internal consistency, structural validity, measurement error, cross-cultural validity, criterion validity, and responsiveness are not relevant for the COSMIN checklist.

Chapter 8 contains a general discussion of the dissertation. In this chapter we concluded that the Delphi technique was suitable to reach consensus on the COSMIN checklist and taxonomy, because the Delphi technique is particularly suitable when there is a lack of knowledge or agreement on a particular subject. Arguments for choices to be made are identified and considered in a Delphi procedure. Because decisions about terminology and definitions are a matter of agreement between experts, and these decisions cannot be investigated empirically, the Delphi technique is useful. The place of each measurement property in the taxonomy is a logical consequence of the chosen definition.
We also concluded in this chapter that the panel involved in this Delphi study was appropriate, because it represented all disciplines involved in HR-PRO measurements. Many of those who agreed to participate remained involved until the process was completed, despite the large burden on their time.
The Delphi technique is sensitive to three forms of bias, i.e. selection bias, subject bias, and bias introduced by the researchers in the interpretation of the findings. To avoid selection bias in this study, we invited a heterogeneous sample of panel members. To avoid subject bias, i.e. respondents are influenced by the opinion of well-known experts, we kept all answers of panel members anonymous. To avoid interpretation bias, we formulated a priori criteria for concluding when consensus was reached, and we tried to be as transparent as possible. This latter was accomplished by asking the panels' opinion, by encouraging them to provide their arguments for their choices, by providing all responses of questions of the previous round in a feedback report, and by explaining our choices.

Furthermore, in this chapter we discussed combining the scores of each box. To be able to give an overall score of the methodological quality of a measurement property, the scores on the COSMIN checklist need to be combined into an overall score for each measurement property. Within the COSMIN initiative, in cooperation with the Working Group on Clinimetrics of the EMGO institute for Health and Care Research, a rating system is being developed to classify examinations of measurement properties into excellent/good/fair/poor methodological quality. This rating system needs to be further evaluated.

In order to be able to determine if the measurement instrument itself is adequate, studies on each of the measurement properties should have good results. Therefore, criteria of adequacy should be developed, preferably consensus-based, and if possible based on empirical evidence for these criteria. Recommendations for using the COSMIN checklist can also be found on the website, www.cosmin.nl.