Patient Related Outcome Measures

Dovepress
open access to scientific and medical research

Open Access Full Text Article

REVIEW

Instrument development and evaluation for patient-related outcomes assessments

This article was published in the following Dove Press journal:
Patient Related Outcome Measures
21 February 2012
Number of times this article has been viewed

Małgorzata Farnik
Władysław Pierzchała

Department of Pneumonology, Silesian University of Medicine, Katowice, Poland

Abstract: Patient-related outcomes measures could provide important information for the current state of the art in medical care and even have an impact on macrodecisions in the health care system. Patient-related outcomes were initially defined as subjective health indicators that allow disability and illness to be assessed, based on patient, caregiver, or physician self-reports. As illness involves psychological and behavioral complex processes of care, a multidisciplinary approach in measuring patient-reported outcomes should be recommended, such as quality of life questionnaires. Patient-related outcomes measures should correspond to specific clinical situations and bring opportunities to improve quality of care. Objective measurements enable quantitative data to be collected and analyzed. Depending on the aim of the research, investigators can use existing methods or develop new tools. This publication presents a methodology for developing patient-related outcomes measures, based on a multistage procedure. The proper definition of specific study objectives and the methodology of instrument development are crucial for successfully transferring the study concept. The model of instrument development is the process of starting from the preliminary phase and includes questionnaire design and scaling, pilot testing (cognitive debriefing), revision of the preliminary version, evaluation of the new tool, and implementation. Validation of the new instrument includes reliability, reproducibility, internal consistency, and responsiveness. The process of designing the new tool should involve a panel of experts, including clinicians, psychologists (preliminary phase), and statisticians (scale development and scoring), and patients (cognitive debriefing). Implementation of a new tool should be followed by evaluation study – assessment of the tool's usefulness in clinical practice. An instrument must show not only the expected methodological properties and performance but also a positive contribution to care. The necessity of implementation of direct patient-reporting methods has been highlighted by both the Food and Drug Administration and the European Medicines Agency.

Keywords: instrument development, patient-related outcomes, evaluation

Introduction

A noticeable trend toward there being a more active role of patients in treatment management has an impact on the development of patient-related outcomes measures. Nowadays, patients' voices support public health strategies, clinical practice, and decision making at an individual level.

Patient-related outcomes were initially defined as subjective health indicators that allow disability and illness to be assessed, based on patient, caregiver, or physician self-reports. A broarder sense of subjective health indicators assumes patients' functioning and ability to perform tasks of daily living. Accordingly, a holistic approach in medicine in most existing patient-related outcomes measures reflects

Correspondence: Małgorzata Farnik
Department of Pneumonology, Silesian University of Medicine, Ul Medyków 14, 40-752 Katowice, Poland
Tel/fax +48 252 38 39
Email pneumo@sum.edu.pl

the patient's perception of the impact of disease and treatment, including multi-item health-related quality of life instruments, single-item measures such as visual analog scale, and other measures, such as daily diaries or treatment adherence.[1,2] According to the Food and Drug Administration, patient-reported outcomes instruments should be recommended as efficacy outcomes in clinical trials, as some illness effects are known only to the patient.[3] The necessity of the implementation of direct patient reporting has been highlighted by the European Medicines Agency. Work on the implementation of new pharmacovigilance legislation (effective July 2012) will bring an opportunity to strengthen the interaction between patients and the health care system.[4]

Objective measurements enable quantitative data to be collected in a standardized manner; thus, the data are internally consistent and coherent for analysis. Questionnaires are the most commonly used instrument for data collection. The information is obtained from the respondent based on a formalized set of questions, which allows respondent self-rating or personal evaluation. Previously developed instruments assessing patient-related outcomes mainly covered the context of functional ability – impairment, disability, and functional handicap. Recently, researchers have focused on clinical judgments based on positive health concepts such as quality of life.[5] As illness involves psychological and behavioral complex processes of care, a multidisciplinary approach in measuring patient-reported outcomes should be recommended, such as quality of life questionnaires. Health-related quality of life is defined as the individual's perception of their position in life and how it is affected by their physical health, psychological state, level of independence, social relationships, and relationship to salient features of their environment.[6]

The model of instrument development

The overriding objective of instrument development is to translate the researcher's information needs into a set of specific questions that respondents are willing and able to answer. The process of instrument development is based on a multistage procedure, which includes the following:

- Preliminary phase: initial questions, reasons for creating the instrument, identification of patients or special groups to which the instrument is addressed, identification of needs, operationalization of variables
- Questionnaire development: questions, scales
- Pilot testing: assessment for feasibility, comprehension, ease of use, usefulness of the instrument, context of the research
- Evaluation: the validation process, including reliability, reproducibility, internal consistency, responsiveness.

Preliminary phase

The aim of the preliminary phase is the identification of patients or a group of respondents, the translation of researchers' needs into variables, and their operationalization. Based on relevant literature, information obtained from the specialist involved in a specific area (such as health care providers), and focus groups, the preliminary version of the instrument is developed. Invitation for focus group discussion should be addressed to identified groups of respondents, covering all the potential spectrum of participants representing different age, gender, and education levels. Focus groups can be particularly helpful in gathering information before developing a survey questionnaire to see what topics are salient to respondents, how people understand a topic area, and how respondents interpret questions. Focus group discussion can also bring up information on how framing a topic or question in different ways might affect responses and whether the topic/question is relevant in their situation. During focus group discussions, the surveyor typically gathers a group of people and asks them questions, both as a group and individually. Focus group moderators may ask specific survey questions, but often focus group questions are less specific and allow participants to provide longer answers and to discuss a topic with others. Starting with broad questions, the moderator typically asks more specific follow-up questions. Focus groups and interviews with specialists provide qualitative data, which is a valuable component of the research process.

There are important limitations of such qualitative study, such as interactions among participants that might have an impact on the opinions expressed by others in the group. The total number of participants is often small, and respondents are not a randomly selected subset of the population. Thus, the moderator should be experienced in focus groups, and the qualitative data must be interpreted with caution.

Questionnaire development

Based on qualitative data provided during preliminary study, the researcher should decide which questions need to be included in the instrument, as well as what type of questions and responding options would be appropriate. It is crucial

to remember that each question should relate directly to the survey questionnaire objectives and should be phrased so that all respondents interpret it the same way. Every respondent should be able to answer every question. If necessary, the respondent should be instructed. Questionnaires should not be long, as brief questionnaires have higher response rates. The questionnaire should start with questions that are easy to answer and avoid asking for identifying information or having the most difficult questions in the beginning of the survey questionnaire. The language used in the questionnaire should be direct and simple; thus, respondents can answer quicker and more accurately. If questions represent domains or subscales, the area of interest should be clearly defined, and the appropriate components should be included in each part of the tool. The preliminary version could contain more questions, and later items could be reduced by choosing the most appropriate questions. If the symptoms and signs are self-assessed, they should be distinguishable from impacts. It is also very important that the recall period is optimal for the concept and population and that the mode of administration is appropriate.

Specific examples

Self-reports and proxy measures

Questionnaires could be addressed to patients or other respondents if the patient is not able to answer the questions (eg, young children).

- Self-report instruments directly measure the patients' perceptions.
- Proxy measures reflect patients' conditions assessed by delegated respondent (caregiver/parent).

Parent proxy measures are available for children aged 2–18 years. However, imperfect agreement between self- and proxy reports (the cross-informant variance) has been documented both in chronic health conditions and in healthy children.[7–9]

The mode of administration

Depending on who is completing the questionnaire, there are the following options:

- Self-completed questionnaires: if the questionnaire is completed by a patient, the respondent should be instructed
- Interviewer-completed questionnaires: if the questionnaire is completed by a trained interviewer.

If the instrument is developed for phone or mail administration, it should be designed differently. Questions in the questionnaire developed for phone administration should not be too long. Also, too many response options could be difficult to remember for the respondent and cause problems when choosing an appropriate answer. For instrument design to be completed by mail, clear examples or repeated instructions would be useful.

Questionnaires are preferred as methods of assessment, but they are also suitable for statistics calculation using a single-item measure that is based on only one question, or a battery if a series of single-item measurements is used to assess the same concept.[10]

Measurement scales

There is a wide variety of scaling methods. The scaling item responses depend on the measured issues and investigators' needs. Most scaling methods used in patient-related outcomes measures are based on dichotomized categories or a continuum. The scales offering a range of choices are preferred, rather than categorical response choices. This allows patients to choose the option of a long continuum agreement rather than to simply agree or disagree with a statement.

Scales are commonly developed based on Thurstone's method, the Likert scale, or Guttman scaling. Thurstone invented three different methods for developing a unidimensional scale: the method of equal-appearing intervals, the method of successive intervals, and the method of paired comparisons.[11] The other option is the Likert scale, which offers respondents items that can be rated 1–5 or 1–7. More recently, Guttman scaling was presented, which is sometimes known as cumulative scaling or scalogram analysis.[12] The purpose of Guttman scaling is to establish a one-dimensional continuum for a measured concept. The key to this scaling method in the analysis is to construct a matrix or table that shows the responses of all the respondents on all of the items. Each scale item has a scale value associated with data obtained from the scalogram analysis. The scale allows degree of disability to be ranked, in respect of a number of particular activities. It is useful if disability progresses steadily from one activity to another (eg, starting from difficulty in walking and then in dressing).[8] Guttman scaling is popular, but it has been criticized for its methods of attributing equal weights to item responses.[13]

The most recommended scale for patient-related outcomes measures is the Likert scale. It is important to use scales that provide the information that is needed and that are appropriate for respondents. Each choice criterion needs to be clearly stated and defined as a single category. The difference between each category should be relevant to the

research objective. When necessary, the respondent should be instructed on how to complete each section and how to mark the answers, to ensure that the survey questionnaire is completed correctly.

There are many options of responses, as follows, of which the fixed responses are most often used:[10]

- Fixed response (quantitative):
  - Single select category (such as yes/no): single select questions are useful for a variety of purposes (eg, prioritizing respondents, asking follow-up questions based on the answer selected, or if filtering the respondents based on their answers is necessary). A few examples of single select questions are preference scales, or "yes/no" options, or selecting the best answer that applies
  - Multiple choice: the multiple select (or multiple choice) scales allow respondents to choose several different responses to a specific question. The scale could be used if the researcher prefers to compare across categories or to focus on particular options. Multiple choice allows recognition of the other respondents' preferences
  - Rating scale or continuum (such as a Likert-type scale): rating scale and continuum are based on a single select format. Matrix questions could be used to group questions with the same set of responses, such as in a Likert scale ("summated scale"). The categories could be defined as separate options and allow respondents to choose one option. Each respondent is asked to rate each item on a response scale (eg, rate each item on a 1–5 response scale where, for example, 1 = strongly disagree, 2 = disagree, 3 = undecided, 4 = agree, 5 = strongly agree). The final score for the respondent on the scale is the sum of their ratings for all of the items. With some scales, if items are reversed in meaning from the overall direction of the scale, it is necessary during calculation to reverse the response value for each of these items before summing the total. These are called reversal items. A continuum scale could be structured as a line with descriptions on both ends: point "0", "none", and "maximum", and "total" as opposite. Respondents are asked to place a mark on the line corresponding to their state. An example of a well-known scale based on a continuum is the visual analog scale
  - Rank ordering: a rank order question allows respondents to state their preferences from a list of items. A ranking order question obtains not only the most preferred items but also the sequence of the remaining items. The relationship or importance between them can be measured. The scaling is useful if the preferences of specific respondents need to be analyzed. The multiple rank order questions could be used to conduct correspondence analysis, in order to quantify and compare different sets of perceptions
- Open-ended questions (qualitative) (limited version and unlimited version): open-ended questions are defined as unstructured or qualitative questions that allow respondents to enter alphanumeric responses. Open-ended questions could be structured as limited and unlimited. Limited text questions require a brief answer; unlimited text questions allow respondents to answer with as much information as they think is necessary. Open-ended questions may be used in qualitative study or for follow-up, in order to collect additional information or for clarification. Unlimited text questions are helpful when they are independent questions that require a higher amount of input from the respondents. The useful method to analyze this data, which is special for the unlimited type of open-ended questions, would be grouping answers into categories. Thus, it would be possible to measure results quantitatively.

Pilot testing

Preliminary version testing is required to ensure the understanding of the newly developed tool. According to recommendations of the MAPI Research Institute (Lyon, France), pilot testing should be based on a sample of around 35 respondents.[14] Participants should be selected randomly. It is optimal if they represent different educational levels and socioeconomic backgrounds. The methodology of cognitive debriefing as the qualitative assessment consists of:[15]

- Comprehension of each question (question intent, meaning of terms)
- Retrieval of memory of relevant information (what types of information do patients need to recall and what types of strategies are used to retrieve information?)
- Decision processes (does the respondent devote sufficient mental effort to answering accurately or does the respondent choose an answer because they think a given answer may be expected from them?)
- The response process (the response options should be clear and allow respondents to choose the appropriate answers)
- General comments (eg, if the questionnaire is considered as being too long).

The aim of the cognitive debriefing is to identify difficult items and confusing questions. This requires explanations

of why; thus, a better version could be proposed. The aim is also to identify whether the interpretation of an item differed between the respondents. It is worthwhile including a question concerning an alternative suggestion of how the question should be asked.

The cognitive debriefing is conducted mainly as the think aloud method or verbal probing techniques. The value of the think aloud method is avoiding interviewer bias and minimal training requirements from the interviewer.[16] The respondent is instructed prior to completion of the instrument to think aloud as he/she answers the questions. This recorded data could be interpreted later in the context of comprehension, the decision process, or other aspects of pilot testing. The technique is used to determine whether the meaning of an item, as intended by the questionnaire developer, is consistent with the respondent's interpretations of that item.[17]

Another option of cognitive debriefing is the verbal probing technique.[18] This is based on face-to-face interviews focused on particular categories of cognitive probes, such as comprehension and interpretation probes; paraphrasing; and general probes, such as whether the patient has found the question to be difficult and whether the scale allows the respondent to answer in the way they would have liked to.

The instrument is revised based on the results of cognitive debriefing.

Evaluation of the new instrument

The next step of instrument development is the validation process. The achievement of standards of validity and reliability requires time and includes rigorous methods of data analysis.

Validity

Validity is the ability of the measure to provide accurate measurements, or is defined as the degree to which an assessment measures what it is supposed to measure. However, different types of validity have emerged, all of which address the issue of degree of confidence that can be placed on the inferences drawn from the scale scores.[19,20]

Construct validity

Construct validity evidence involves empirical and theoretical support for the interpretation of the hypothetical construct, representing mainly psychology or sociology. It is important if the theories attempt to explain behaviors and attitudes. Construct validity includes statistical analyses of the internal structure of the instrument and the relationships between responses to different test items or measures of other constructs. Many predictors could be made from one construct. The validity of the measure could be problematic if the predictors made on the basis of theory are not confirmed. This type of validity is divided into convergent and discriminant validity, which involves the extent to which the instrument is related to other variables and measures of the same construct.[10,20]

Convergent validity refers to the degree to which a measure is correlated with other measures that it is theoretically predicted to correlate with. Thus, convergent validity requires that the instrument should correlate with other measures of this construct.

Discriminant validity describes the degree to which the variables do not correlate with other dissimilar variables that theoretically they should not correlate with.

Content validity

Content validity describes how the components of the instrument cover all of the attribute to be measured and if the number of items in each area or subscale reflects its importance. Content validity evidence involves the degree to which the content of the test matches a content domain associated with the construct, and is a nonstatistical type of validity that involves "the systematic examination of the test content to determine whether it covers a representative sample of the behavior domain to be measured."[10,21]

Face validity

Face validity is one of the forms of content validity that estimate whether an instrument appears to measure a certain criterion. It does not guarantee that the test actually measures phenomena in that domain, but rather indicates whether items appear to measure the variables they claim to measure. Face validity relates to whether or not an instrument appears to be a good measure. This judgment is made on the "face" of the measure (thus named as "face validity") and can also be judged by amateurs.[10]

Criterion validity

Criterion validity indicates whether the variable can be measured with accuracy. The definition of criterion validity is the correlation of a scale with some other measure of the trait study. It thus involves the correlation between the instrument and a criterion variable, which is taken as representative of the construct. Criterion validity is divided into two categories: concurrent or predictive validity. If the test data and criterion data are collected at the same time, this is referred to as concurrent validity evidence. Thus, both scales should be administered

at the same time. If the test data are collected first in order to predict criterion data collected at a later point in time, then this is referred to as predictive validity evidence. It refers to the degree to which the variable can predict (or correlate with) other measures of the same construct that are measured at some time in the future. A high correlation would provide evidence for predictive validity of the instrument.[10,19,20]

Known groups validity

Known groups validity is a form of construct validation in which the validity is determined by the degree to which an instrument can demonstrate different scores for groups known to vary on the variables being measured.[19,20]

Longitudinal validity

Longitudinal validity is the extent to which changes in one measure will correlate with changes in another measure.[19,20]

Reliability

The new instrument requires testing for reliability. Reliable measure means that it is stable or consistent and produces similar results when administered repeatedly, when there is no evidence of change. There are many variations on the measurement of reliability, including internal consistency, inter-rater agreement, intrarater agreement, test-retest, and sensitivity to change.

Internal consistency

Testing for homogeneity of the measurement is an important procedure assessing the reliability of the instrument. Internal consistency is defined as the correlation between the items in the scale or within each scale domain, or correlation between the items and the total score. If the correlation concerns two halves of the scale, where the scale can be divided into equivalent parts, the split half reliability could be assessed. Internal consistency is measured by applying Cronbach's α coefficient. The calculation is based on an average correlation among the items and the number of items in the instrument. Thus, the coefficient reaches values between 0 and 1. The expected level of Cronbach's α is over 0.5.[10,22]

Test-retest reliability

To assess test-retest reliability, the instrument is administered to the same population on two occasions (stable over the interval between assessments), and the two scores are assessed for consistency. The results could be influenced by the possibility of practice effects, which can artificially inflate the estimate of reliability.[19,20]

Inter-rater reliability

Inter-rater reliability determines the extent to which two or more raters obtain the same result when using the same instrument.

Intrarater reliability

Intrarater reliability agreement is the reliability of the same rater's scores of the same subjects on different occasions.

Sensitivity to change

Sensitivity to change is defined by the instrument's responsiveness to detecting the change. It requires correlating its scores with other measures that reflect any anticipated changes.

Future directions

During the past years, the substantial increase in pediatric clinical trials has been noted as the result of legislation changes. The Food and Drug Administration requires pediatric studies if the product is likely to be used by a considerable number of children as the new treatment option. As evaluation of patient-related outcomes is recommended with assessment of the efficacy of treatment, there is a need to implement specific measures. Despite of the possibility to self-report on matters pertaining own health and well- being the reliability of such assessment by pediatric patients is discussed.[23] Patient reports or proxy measures have been documented in many clinical trials. Evaluation conducted including children with chronic conditions has proved that children as young as 5 years old are reliable when reporting intensity of pain and discomfort.[24] There is a need for studies assessing the agreement of self-reports and proxy measures. Another important issue is defining when young children can reliably and validly self-report outcomes, and which measures should be recommended in the context of development psychology and verbalization skills.

Conclusion

Patient-related outcomes measures should correspond to specific clinical situations and bring the opportunity to improve quality of care. Implementation of new tools should be followed by an evaluation study and assessment of its usefulness in clinical practice.[25] An instrument must not only show expected methodological properties and performance but also offer a positive contribution to patient care.[26] The instrument should be supported by a brief, comprehensive bibliography of the most important references.

For successful transferring of the concept of research to new instrument development and implementation, investigators should start with a strong definition of specific study objectives and follow a methodology of instrument development. The process of designing a new tool should involve a panel of experts, including clinicians, psychologists (preliminary phase), and statisticians (scale development, scoring), as well as patients (cognitive debriefing). Patient-related outcomes measures could provide important data for the current state of the art in medical care and even have an impact on macrodecisions.

Disclosure

The authors declare no conflict of interest.

References

1. Acquardo C, Berzon R, Dubois D. Incorporating the patient's perspective into a drug development and communication. An ad hoc task force report of the Patient-Reported Outcomes (PRO) harmonization group meeting at the Food and Drug Administration, Feb 2001; *Value in Health*. 2003;6:522–531.
2. Willke RJ, Burke LB, Erickson P. Measuring treatment impact: a review of patient-reported outcomes and other efficacy endpoints in approved product labels. *Contr Clin Trials*. 2004;25:535–552.
3. Center for Drug Evaluation and Research. US Food and Drug Administration guidance for industry. Patient-reported outcome measures: use in medical product development to support labeling claims. Rockville, MD: Food and Drug Administration; 2006.
4. European Medicines Agency. Fourth report on the progress of the interaction with patients' and consumers' organizations (2010) and results/analysis of the degree of satisfaction of patients and consumers involved in EMA activities during 2010. EMA/632696/2011.
5. World Health Organization. International classification of impairments, disability and handicaps. Geneva, Switzerland: World Health Organization; 1980.
6. World Health Organization. Measurement of quality of life in children. Geneva, Switzerland: Division of Mental Health, World Health Organization; 1993.
7. Koot HM, Wallander JL, editors. Quality of life in child and adolescent illness: concepts, methods and findings. East Essex, UK: Brunner-Routledge; 2001.
8. Varni JW, Burwinkle TM, Seid M, Skarr D. The PedsQL 4.0 as a pediatric population health measure: feasibility, reliability and validity. *Amb Ped*. 2003;3:329–341.
9. Farnik M, Pierzchała W, Brożek G, Zejda JE, Skrzypek M. Quality of life protocol in early asthma diagnosis in children. *Ped Pulm*. 2010;45:1095–1102.
10. Bowling A. Measuring health. A review of quality of life measurement scales. Buckingham, UK: Open University Press; 1998.
11. Streiner DE, Norman GR. Health measurement scales: a practical guide to their development and use. Oxford, UK: Oxford University Press; 1989.
12. Williams RGA, Johnston M, Willis M. Disability: a model and measurement technique. *Brit J Prevent Soc Med*. 1989;30:71–78.
13. Skinner DE, Yett DE. Debility index for long-term care patients. In Berg RL, editor. Health status indexes. Chicago, IL: Hospital Research and Education Trust; 1976.
14. The MAPI Linguistic Validation Process. Available from: http://www.mapi-research-inst.com. Accessed January 17, 2012.
15. Campanelli P. Testing survey questions: new directions in cognitive debriefing. *Bull Methodol Soc*. 1997;55:5–17.
16. Davidson GC, Vogel RS, Coffman SG. Think-aloud approaches to cognitive assessment and the articulated thoughts in simulated situations paradigm. *J Consult Clin Psychol*. 1997;65(6):950–958.
17. Willis GB. Cognitive interviewing: a tool for improving questionnaire design. Thousand Oaks, CA: Sage; 2005.
18. Willis GB, Royston P, Bercini D. The use of verbal report methods in the development and testing of survey questionnaires applied. *Cog Psych*. 1991;251–267.
19. American Educational Research Association, Psychological Association, and National Council on Measurement in Education. Standards for educational and psychological testing. Washington, DC: American Educational Research Association; 1999.
20. Lang T, Secic M. How to report statistic in medicine. Annotated guidelines for authors, editors and reviewers. 2nd ed. Philadelphia, PA: American College of Physicians; 2006.
21. Urbina S, Anastazi A. Psychological testing. Upper Saddle River, NJ: Prentice Hall; 1997.
22. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometr*. 1951;22:293–296.
23. Clarke SA, Eiser C. The measurement of health-related quality of life in pediatric clinical trials: a systematic review. *Health Qual Life Out*. 2004;2(66):1–5.
24. Varni JW, Bernstein BH. Evaluation and management of pain in children with rheumatic diseases. *Rheu Dis Clin North Am*. 1991;17:985–1000.
25. Nelson EC, Berwick DM. The measurement of health status in clinical practice. *Med Care*. 1989;27 Suppl 3:77–90.
26. Klauser AG, Schindlbeck NE, Muller, Lissner SA. Symptoms in gastroesophageal reflux disease. *Lancet*. 1990;335:205–208.