Volume 11 • Number 4 • 2008
VALUE IN HEALTH

Development of EMPRO: A Tool for the Standardized Assessment of Patient-Reported Outcome Measures

Jose M. Valderas, MD, PhD, MPH,[1,2,3] Montse Ferrer, MD, PhD, MPH,[2,3] Joan Mendívil, MD,[2,4] Olatz Garin, MPH,[2,3] Luis Rajmil, MD, PhD,[2,3,4,5] Michael Herdman, MSc,[2,3] Jordi Alonso, MD, PhD,[2,3,6] on behalf of the Scientific Committee on "Patient-Reported Outcomes" of the IRYSS Network*

[1]National Primary Care Research and Development Center, University of Manchester, Manchester, UK; [2]Health Services Research Unit, Institut Municipal d'Investigació Mèdica (IMIM-Hospital del Mar), Barcelona, Spain; [3]CIBER en Epidemiologia y Salud Pública (CIBERESP), Spain; [4]Agència de Salut Pública de Barcelona, Barcelona, Spain; [5]Agència d'Avaluació de Tecnologies i Recerca Mèdica (AATRM) de Catalunya, Barcelona, Spain; [6]Universitat Pompeu Fabra, Barcelona, Spain

ABSTRACT

Objective: This study was aimed to develop a tool for the standardized assessment of patient-reported outcomes (PROs) to assist the choice of instruments.

Methods: An expert panel adapted the eight attributes proposed by the Medical Outcomes Trust as evaluation review criteria, created items to evaluate them, and included a response scale for each item. A pilot test was designed to test the new tool's feasibility and to obtain preliminary information concerning its psychometric properties. The Spanish versions of five measures were selected for assessment: the SF-36 Health Survey, the Nottingham Health Profile, the COOP-WONCA charts, the EuroQol-5D, and the Quality of Life Questionnaire EORTC-QLQ-C30. We assessed the new tool's reliability (Cronbach's alpha and intraclass correlation coefficient [ICC]) and construct validity.

Results: The new EMPRO (Evaluating the Measurement of Patient-Reported Outcomes) tool has 39 items covering eight key attributes: conceptual and measurement model, reliability, validity, responsiveness, interpretability, burden, alternative modes of administration, and cross-cultural and linguistic adaptations. Internal consistency was high ($\alpha = 0.95$) as was interrater concordance (ICC: 0.87–0.94). Positive associations consistent with a priori hypotheses were observed between EMPRO attribute scores and the number of articles identified for the measures, the years elapsed since the publication of the first article, and the number of citations.

Conclusion: A new tool for the standardized assessment of PRO measures is available. It has shown good preliminary reliability and validity and should be a useful aid to investigators who need to choose between alternative measures. Further assessment of the tool is necessary.

Keywords: assessment, health-related quality of life, patient-reported outcomes, psychometric properties.

Introduction

The search for health outcome measures that incorporate the patient's perspective has led to the production of an increasing number of "patient-reported outcome" (PRO) measures [1,2]. This has been in part due to new models of care with greater patient involvement in decision-making, as well as the recognition that these measures can independently predict the use of health services, such as hospitalization, primary care, health-care needs, and even mortality [3,4]. They also elicit information which has found to be complementary to commonly used clinical measures [5]. Recent estimates, based solely on publications in English, place the number of measures at 1275 [6]. This poses at least two problems: 1) how to identify questionnaires which are available for a specific use (e.g., a study of functional status in patients with asthma); and 2) how to choose the most appropriate measure from among those available. Some initiatives have addressed the first of these problems by compiling catalogues of questionnaires in different formats (including books and electronic databases) [7–12].

As regards the second question, criteria are needed to identify the strengths and weaknesses of the PRO measures. These criteria should cover the conceptual and theoretical model on which the measure is based,

Address correspondence to: Jordi Alonso, Health Services Research Unit, Institut Municipal d'Investigació Mèdica (IMIM-IMAS), Dr Aiguader, 88. Barcelona-08003, Spain. E-mail: jalonso@imim.es

10.1111/j.1524-4733.2007.00309.x

*The Scientific Committee on "Patient-Reported Outcomes" of the IRYSS Network: J. Alonso (URSS-IMIM), B. Bolívar (FJGG), A. Domingo-Salvany (URSS-IMIM), A. Escobar (UIPV), M. Ferrer (URSS-IMIM), M. J. Forjaz (ISCIII), J. M. Haro (FSJ-SSM), J. del Llano (FGC), J. Mendivil (URSS-IMIM), P. Martínez (ISCIII), G. Permanyer (HVH), L. Rajmil (AATRM), P. Rebollo (HCA), A. Ribera (HVH), R. Santed (URSS-IMIM), V. Serra-Sutton (AATRM), and J. M. Valderas (URSS-IMIM).

  1098-3015/08/700 700–708

as well as its psychometric properties (reliability, validity, or sensitivity to change). Several attempts have been made to systemize review criteria for PRO measures, all of which are inspired by the seminal work of McDowell et al. [7]. Several attempts have been made to systematize review criteria for PRO measures. The GraQol Index was the first attempt to develop a tool that generated a global score [13]. Although promising, the authors recognized several limitations in the evaluation of specific attributes and criteria selected which limited its use [13]. The majority of other initiatives have not transcended the use they were originally designed for with one possible exception, namely the proposal by the Scientific Advisory Committee of the Medical Outcomes Trust (MOT). The MOT is a nonprofit organization in the United States, which aims to provide quality control guidelines for the development of measures that incorporate the patient's perspective in the field of health service research [14]. Based initially on sound theoretical grounds, the MOT initiative was recently expanded to include an explicit guide as to how each criterion should be met [15].

The aims of the present study were to apply these recent advances to develop a tool for the standardized assessment of PRO measures and to obtain preliminary information on its validity and reliability.

Methods

The Spanish Cooperative Investigation Network for Health and Health Service Outcomes Research (Red-IRYSS) was established in 2002 and had as one of its main goals the promotion of health-related quality of life measures for use with Spanish-speaking populations [16]. As it was considered that one way of achieving this aim would be to have a tool available for the standardized assessment of PRO measures, members of the Red IRYSS reviewed earlier initiatives to produce such a tool. After review, the updated version of the MOT proposal was selected as providing the most comprehensive and explicit set of recommendations available [15]. This proposal is the broadest in scope in terms of the attributes assessed, as it includes the psychometric characteristics of the questionnaires as well as the administration procedures; it also provides the most thorough specification of criteria that need to be met by PRO measures.

Development

As a first step in the development of the new tool, a panel of four experts was nominated, based on their substantive experience in the development, assessment, and use of PRO measures (see Acknowledgments for the members of the panel). The panel operatively defined the quality of a PRO instrument as the "degree of confidence that all possible bias has been minimized and that the information about the process which led to its development and evaluation is clear and accessible."

The content of the new tool was generated by converting each of the specific criteria in the MOT proposal into individual items. The expert panel aimed to respect the original wording as far as possible and to maximize standardization of the assessment-incorporated quantitative criteria based on the literature whenever possible [14–18].

The format of the instrument (response options and structure) was based on the AGREE (Appraisal of Guidelines Research and Evaluation) instrument, a tool originally conceived for the appraisal of clinical practice guidelines [19]. This instrument has been shown to have excellent properties, and it has been widely disseminated [19–22].

The resulting tool was then circulated to other researchers for their views on the clarity of its content, comprehensiveness, and ease of use (see Acknowledgments). A final version was obtained which incorporated their suggestions, and a user's manual was produced.

Assessment Process

Once the tool was created, the expert panel defined the process to be followed when evaluating a PRO measure, consisting on the assessment by several reviewers of standardized information retrieved for each PRO measure.

Evaluation of Psychometric Properties

A pilot study was performed to test the new tool's feasibility of use and measurement properties (score distributions, reliability, and validity). The measures assessed were the Spanish versions of five well-known and widely used PRO instruments. These included four commonly used generic questionnaires [6] (the SF-36 Health Survey [23,24], the Nottingham Health Profile [25,26], the COOP-WONCA Charts [27,28], and the EuroQol-5D [29,30]), and an instrument used in patients with cancer (the Quality of Life Questionnaire EORTC-QLQ-C30 [31,32]).

Twenty reviewers were invited to participate based on their experience in using and evaluating PRO measures. Only one reviewer declined to participate. All but four reviewers were members of the Scientific Committee on "Patient-Reported Outcomes" of the IRYSS Network. The others were external researchers without any formal link to the project. Five panels were formed to review the selected instruments: four panels had four reviewers each and one panel had three reviewers. The four member panels each included one of the external reviewers. Each panel was assigned one of the selected instruments to assess. Within each panel, instruments were assessed independently by the reviewers; reviewers had not been involved in the development or adaptation of the instrument assigned.

In order to complete the assessment, each reviewer was provided with the new tool, together with the following documents:

- The original and Spanish versions of the instrument to be assessed;
- The user manual for the Spanish version;
- Related full-text publications (articles, conference papers, information provided by the authors and other documents), which provided information concerning the adaptation into Spanish;
- An article describing the conceptual model of the original questionnaire (not the Spanish version);
- Publications on the validation of the Spanish version and other information related to psychometric properties; and
- A list of all the references identified related to applications of the questionnaire (with summaries when available).

All these documents were identified through a systematic review performed for each questionnaire [33]. Selected databases included PubMed [34,35], ISI Web of Knowledge [36], ProQOLID [37], as well as the Iberoamerican bibliographic databases LILACS [38] and SCIELO [39], and local databases of Spanish biomedical literature, including databases of doctoral theses [40–43]. Specific search definitions for each database are available from the authors upon request. Nonelectronic sources were also searched [6,8–12,44]. All searches were performed between April and June 2004.

The documentation was made available to reviewers both on a CD-ROM and on a secure web site with password-protected access. Materials also included written instructions on the assessment procedure and reviewers were asked to complete the review within 30 days.

Analytic Strategy

In order to assess the feasibility of EMPRO, the proportion of assessments that were returned within the established time period and the number of missing values per assessment was determined. Floor and ceiling effects were calculated for each attribute as the percentage of assessments with the minimum (floor) and maximum (ceiling) scores [45].

Reliability was assessed by examining internal consistency and reproducibility [46]. Internal consistency was calculated using Cronbach's alpha coefficient for each attribute in each panel, and a global estimate in each panel was also obtained (median). Reproducibility was measured as interrater concordance on attribute scores and was assessed using the intraclass correlation coefficient (ICC) for random effects in each panel (two-way model, absolute agreement) [47]. In order to determine whether being an external reviewer affected the assessment, we performed face-to-face comparisons for all reviewer pairs in each panel including external reviewers (ICC). For all reliability measurements, values greater than 0.7 were considered acceptable, and values greater than 0.9 were considered highly satisfactory [17].

A priori hypotheses were formulated with regard to construct validity. We hypothesized that instrument quality would be related to the quality of the publications featuring the instrument. The rationale was that studies describing well-developed, thoroughly tested instruments would have a greater chance of being published in higher-quality publications. The quality of the related publications was defined in terms of selected bibliometric characteristics, namely: 1) absolute number of publications; and 2) citations referring to the first publication (ISI Journal Citation Reports) [48]. The time the instrument had been available (number of years elapsed since the publication of the first article) was also taken into account to determine its effect on the assessment. The Spearman correlation coefficient was used to measure the association between these variables and the attribute scores [49]. Given the small sample of measures ($n = 5$), emphasis was placed on direction of the associations observed rather than statistical significance. All analyses were carried out using the R software (R Foundation for Statistical Computing, Vienna, Austria) [50].

Results

The EMPRO Tool

The new tool consisted of 39 items organized in eight attributes: Conceptual and measurement model (7 items); Reliability (8); Validity (6); Responsiveness (3); Interpretability (3); Administration burden (7); Alternative modes of administration (2); and Cross-cultural and linguistic adaptations (3). Each item consisted of a statement together with a short text to help in its interpretation and application. Reviewers expressed their degree of agreement with the statement on an ordinal Likert-type response scale: "Strongly agree" (4), "Agree" (3), "Disagree" (2), and "Strongly disagree" (1). Space was provided on each item for comments and references to relevant publications or documents.

At the end of the instrument, reviewers were requested to provide an overall recommendation for the measure and provide a rationale for their recommendation. Possible choices were: "Strongly recommended," "Recommended with provisos or alterations," "Would not recommend," and "Unsure."

Scores were assigned based on attribute scores, as well as on the overall degree of recommendation. Attribute scores were calculated as the mean of the responses to all items for that attribute, with a lineal

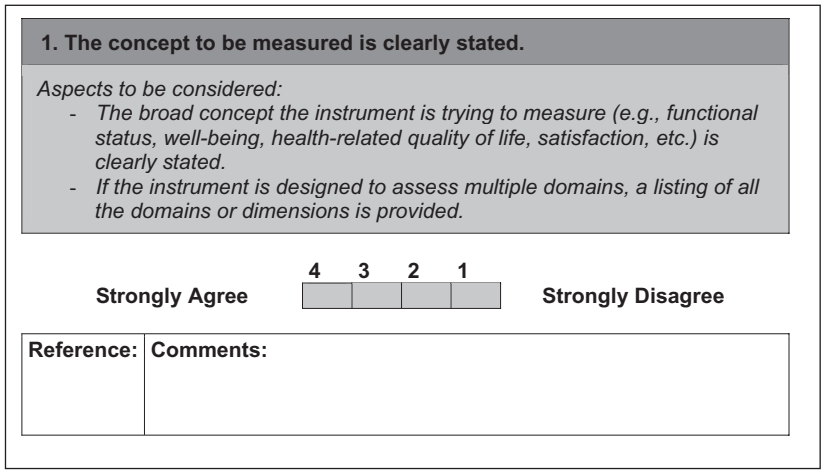

Figure 1 Sample item from the EMPRO tool.

transformation to obtain the scores on a scale from 0 (minimum) to 100 (maximum).

The new tool was named EMPRO (Evaluating the Measurement of Patient-Reported Outcomes). Examples of item structure and overall ratings are shown in Figures 1 and 2, respectively. Full content is available in Appendix 1 (Supplementary material).

Pilot Test

Materials for the evaluation of selected measures. The questionnaire-specific systematic reviews identified a variable number of articles, ranging from 6 to 18. All of the articles detailing the conceptual model of the original version of the questionnaire (n = 5, one per questionnaire) and most of those reporting the adaptation into Spanish were published in international journals (median proportion of international publications = 67%, range from 20% to 80%). The number of citations for the first reference reporting the cultural adaptation ranged from 133 to 0. A user manual was identified for all of the instruments.

Feasibility. All assessments were carried out within the required time period (1 month) and there were no missing data. Nine reviewers (47.4%) completed a questionnaire on the documentation provided, and all found the documentation "Relevant" or "Very relevant" and the selection of materials "Quite appropriate" or "Very appropriate." One reviewer considered the information "Insufficient."

Attribute scores for individual reviewers ranged from 22 to 100 (median scores ranged from 25 to 100) (Table 1). No floor effect was therefore observed. Three attributes (Reliablity, Validity, and Responsiveness) did not receive the maximum score (100) on any of their individual items. A ceiling effect was observed for the attributes of cross-cultural and linguistic adaptations (36.8%) and Interpretability (15.8%).

Reliability and validity. As regards internal consistency, Cronbach alpha coefficients were above 0.7 for all attributes (Table 1). Internal consistency was very high when calculated for all EMPRO items as a whole (median $\alpha = 0.95$). The degree of agreement was

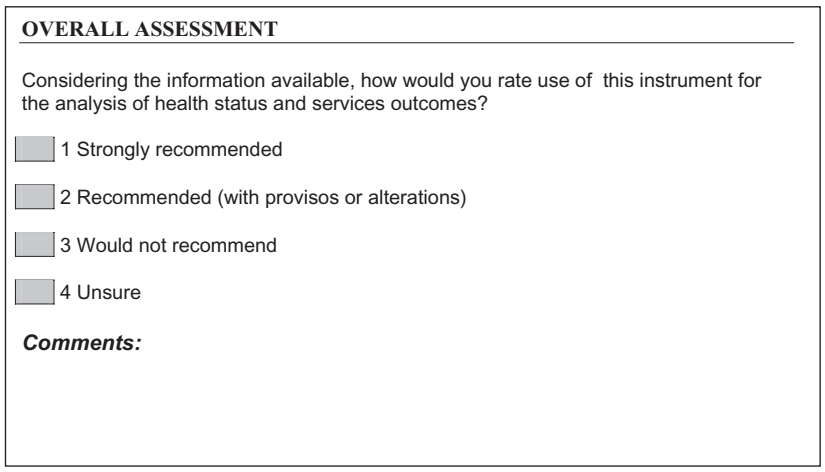

Figure 2 Overall degree of recommendation.

Table I EMPRO attributes' scores (median and range) across measures, and alpha coefficients for internal consistency (range when item deleted)

EMPRO attributes (items)	EuroQol-5D (n = 4; ICC: 0.94 [0.91–0.97])	NHP (n = 4; ICC: 0.91 [0.86–0.94])	MOS SF-36 (n = 4; ICC: 0.92 [0.88–0.95])	EORTC QLQ-30 (n = 4; ICC: 0.93 [0.87–0.96])	COOP charts (n = 3; ICC: 0.87 [0.75–0.93])	Cronbach's alpha
Conceptual and measurement model (7)	84 (79–100)	88 (75–89)*	84 (82–89)	66 (61–93)	46 (39–89)†	0.81 (0.75–0.84)
Reliability (8)	50 (44–59)	80 (59–97)*	64 (59–75)	47 (38–69)†	50 (50–72)	0.75 (0.69–0.77)
Validity (6)	90 (42–92)*	85 (79–96)	75 (58–92)	63 (50–79)	58 (50–75)†	0.79 (0.68–0.80)
Responsiveness (3)	54 (50–92)	75 (33–83)*	63 (33–75)	63 (33–83)	42 (22–67)†	0.74 (0.33–0.93)
Interpretability (3)	92 (83–100)*	88 (58–100)	75 (58–100)	42 (33–67)	33 (25–58)†	0.83 (0.61–0.91)
Administrative burden (7)	96 (86–100)*	70 (46–96)	68 (32–79)	59 (50–61)†	68 (61–71)	0.83 (0.78–0.86)
Alternative modes of administration (2)	75 (38–100)*	50 (25–75)	44 (25–63)	38 (25–63)	25 (25–50)†	0.80 (—)
Cross-cultural and linguistic adaptations (3)	100 (83–100)*	92 (58–100)	88 (50–100)	46 (42–83)†	83 (67–83)	0.71 (0.32–0.79)
Overall recommendation	Strongly recommended	Strongly recommended	Strongly recommended	Strongly recommended	Recommended	Not applicable

*PRO measure with highest median score for this attribute.
†PRO measure with lowest median score for this attribute.
EMPRO, Evaluating the Measurement of Patient-Reported Outcomes; ICC, intraclass correlation coefficient (two-way random-effects model, absolute agreement) (95% confidence interval); MOS SF-36, the MOS 36-Item Short-Form Health Survey; NHP, Nottingham Health Profile.

similar between members of the scientific committee and external assessors and members of the scientific committee (data not shown). We also explored the effect of the availability of information on agreement (ICC): no correlation was found between the degree of concordance and the number of articles published (Spearman Rho = –0.05, $P = 0.93$), but a correlation was found between the degree of concordance and years elapsed since the publication of the first article about the adapted version (Spearman Rho = 0.89, $P < 0.05$).

Positive associations consistent with our a priori hypotheses were observed between EMPRO attribute scores and the number of articles identified for each questionnaires and the number of citations (Table 2). The only exceptions were correlation coefficients close to 0 between scores on the attribute of "Responsiveness" and number of articles, and between "Administrative Burden" and the number of citations. The overall recommendations also showed the expected associations. All correlations observed between years elapsed since the publication of the first article and EMPRO scores were positive.

Discussion

A new tool has been developed, based on earlier explicit recommendations, to aid in the standardized assessment of PRO measures. It has been successfully used to evaluate the Spanish versions of five well-known PRO measures, and the results of the pilot study suggest that it is feasible to use and provide preliminary evidence of its psychometric properties.

The study had a number of limitations. First, the results are based on a small number of reviewers and instruments. Although we did not observe any substantial departure of the assumptions for the calculation of the indices and statistical tests in our study, further replication of our methods in larger samples is necessary. Second, the measures used to evaluate EMPRO's psychometric properties were Spanish versions. It is likely that the retrieval of information for the original instruments would have resulted in an increased number of publications. Nevertheless, it is possible that EMPRO's performance would improve with an increase in the amount and quality of information available for the measures assessed. Third, the assessment of the construct validity for this new tool is still tentative. The bibliometric criteria and their hypothesized associations with the scores were proxies of scientific quality, and they may therefore not represent the quality of a particular manuscript. Fourth, although we asked some researchers about their views on clarity, comprehensiveness, relevance, and ease of use of the tool, we did not perform a formal cognitive pretest before this pilot testing. Nonetheless, agreement regarding the quality of the measures assessed

Table 2 Construct validity: correlations (Spearman Rho) between EMPRO median attribute scores and selected bibliometric characteristics of the measures

EMPRO attributes	No. of articles[†]	Years elapsed[‡]	JCR citations[‡]
Conceptual and measurement model	0.79	0.45	0.87
Reliability	0.26	0.45	0.72
Validity	0.89*	0.80	0.63
Responsiveness	0.05	0.89*	0.36
Interpretability	0.97*	0.57	0.63
Administrative burden	0.53	0.22	–0.15
Alternative modes of administration	0.97*	0.57	0.58
Cross-cultural and linguistic adaptations	0.97*	0.34	0.63
Overall recommendation	0.56	0.79	0.54

*$P < 0.05$.
[†]Number of articles identified concerning the development/adaptation of the Spanish version of the questionnaire.
[‡]Years elapsed and citations for the first publication only.
EMPRO, Evaluating the Measurement of Patient-Reported Outcomes; JCR, Journal Citation Reports.

was high among reviewers, thereby suggesting that the items were actually understood in a similar way by the respondents.

Characteristics of the EMPRO Tool

The values obtained for internal consistency can be considered satisfactory, with all attributes reaching 0.7 and four of them (50%) with coefficients of 0.8 or higher. As reliability depends on the number of items and observations (in part), further studies should include a larger number of measures, and alternative approaches using modern test theory could be applied [51]. The observed level of interrater agreement can also be considered very satisfactory.

Our data also suggest that the EMPRO tool is valid. The method used to develop the instrument supports its content validity, and as a matter of fact the MOT proposal has been very recently used in a similar initiative [52]. The expected associations were observed between EMPRO scores and the variables proposed, supporting the construct validity of the tool. Nevertheless, these relations are also consistent with the hypothesis that EMPRO scores depend on the quantity and quality of the information provided for the assessment. The correlation found between the scores and the information provided (quantity and quality) stresses the need for the information to be identified by means of standardized procedures (systematic review) and be subsequently organized according to explicit criteria.

It was not possible to evaluate the responsiveness of the new tool in the present study. This could be linked to the availability of new information concerning questionnaires that have previously been assessed, based on either new articles, conference papers, or information provided directly by the author.

Applicability

The availability of an instrument to assess the psychometric characteristics and ease of use of PRO measures entails a significant advantage for diverse fields of application, such as clinical, administrative, and research applications. The diversity of existing questionnaires and their relatively recent development make it difficult to identify the most appropriate one for a given use such as monitoring patients in clinical practice [53], identifying population preferences, or choosing the most appropriate questionnaire to evaluate the effect of a therapeutic intervention in a clinical trial. The lack of success of previously developed assessment tools is due largely to the fact that they were developed for use within specific contexts, such as in compilations of PRO questionnaires, and therefore, with rare exceptions, have not been disseminated independently. The instrument we propose has the advantage that it has been developed based on the updated recommendations of experts and that it explicitly specifies review criteria. Researchers will also find in EMPRO a powerful tool in their practice, particularly when considering recent FDA guidance which has established an iterative PRO Instrument Development and Modification Process that explicitly recognizes the assessment of measurement properties as a key step in the refinement of the instruments [54]. Specifically, each of the four measurement properties that the FDA guidance recommend to review (reliability, validity, ability to detect change, and choice of methods for interpretation) corresponds to one of the EMPRO attributes.

Four considerations are worth making for future users of the tool. First, the EMPRO tool relies completely on information obtained in studies in which the instruments are assessed within a specific application. Second, overall ratings and attribute scores are conceptually different measures which are not intended to be interchangeable. The latter are based on the application of explicit criteria, while the overall ratings involve making a subjective recommendation for the instrument given the specific purposes (concept and populations) it was designed for. Both measures provide relevant complementary information that should be considered together when comparing ques-

tionnaires. Third, given the lack of experience with this newly available tool, explicit guides to interpretation have not yet been developed. Available scores for five of the most widely used instruments might prove a useful starting point. Fourth, the choice of any measure will depend on the intended use. In this context, the assessment with EMPRO of several competing PRO questionnaires could help potential users to identify those meeting minimal criteria to be included in the selection. Furthermore, some attributes could assist the choice of the most suitable measure for a specific application. For example, information on responsiveness will be particularly relevant for longitudinal studies designed to evaluate therapeutic interventions, while administrative burden might be of particular relevance if measures are to be used in quality programs within health-care delivery systems.

Finally, the application of EMPRO assumes expert knowledge in the use and assessment of PRO measures. Although it is true that the instrument includes a basic guide to the main concepts, reviewers need a minimum level of knowledge to guarantee interrater concordance. A set of training materials to prepare people who will use EMPRO is currently being developed. Moreover, given the learning curve, it is not unreasonable to expect that interrater concordance would increase if the same reviewers reassessed the same instruments using EMPRO. In the same way, the minimum amount of information needed for each instrument is difficult to establish.

Conclusions

The EMPRO is a new tool that is available to assess the measurement properties and ease of use of PRO measures. Preliminary testing supports its validity and reliability, but replication of our observations is needed, using different PRO measures and reviewers. If these promising results are confirmed, EMPRO should facilitate the selection of the most appropriate PRO measure among competing instruments.

We would like to thank Alfredo Larrea (Hospital Virgen del Camino, Pamplona), Miguel Ruiz (Universidad Autónoma de Madrid, Madrid), Michael Herdman (Institut Municipal d'Investigació Mèdica, Barcelona), and Manuel Salamero (Universidad de Barcelona, Barcelona) for their participation as external reviewers in the pilot test and Vladimir Pizarro (Health Services Research Unit, Institut Municpal d'Investigacio Medica) for his participation in the field work of the pilot test. The panel of experts was composed by: J. Alonso, M. Ferrer, J. M. Valderas (Health Services Research Unit, Institut Municipal d'Investigacio Medica), and L. Rajmil (Catalan Agency for Health Technology and Research Assessment). Additional support in the development of the tool was provided by O. Garin (Health Services Research Unit, Institut Municpal d'Investigacio Medica), A. Ribera (Epidemiology Unit, Department of Cardiology, Vall d'Hebron Hospital), and V. Serra-Sutton (Catalan Agency for Health Technology and Research Assessment).

Source of financial support: This research has been funded by the Instituto de Salud Carlos III (Red de Investigación Cooperativa IRYSS, G03/202 and CM 300118), and Instituto de Salud Carlos III (Exptes. PI06/90483 and ETE06/450104884).

Supplementary material for this article can be found at: http://www.ispor.org/publications/value/ViHsupplementary.asp

References

1 Willke RJ, Burke LB, Erickson P. Measuring treatment impact: a review of patient-reported outcomes and other efficacy endpoints in approved product labels. Control Clin Trials 2004;25:535–52.
2 Acquadro C, Berzon R, Dubois D, et al. for the PRO Harmonization Group. Incorporating the patient's perspective into 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, February 16, 2001. Value Health 2003;6:522–31.
3 Connelly JE, Philbrick JT, Smith GR Jr, et al. Health perceptions of primary care patients and the influence on health care utilization. Med Care 1989;27(3 Suppl.):S99–109.
4 Idler EL, Angel RJ. Self-rated health and mortality in the NHANES-I Epidemiologic Follow-up Study. Am J Public Health 1990;80:446–52.
5 Alonso J. La medida de la calidad de vida relacionada con la salud en la investigación y la práctica clínica. Gac Sanit 2000;14:163–7.
6 Garratt A, Schmidt L, Mackintosh A, Fitzpatrick R. Quality of life measurement: bibliographic study of patient assessed health outcome measures. BMJ 2002;324:1417.
7 McDowell I, Newell C. Measuring Health. A Guide to Rating Scales and Questionnaires. New York: Oxford University Press, 1996.
8 Bowling A. Measuring Disease. A Review of Disease-Specific Quality of Life Measurement Scales (2nd ed.). Buckingham: Open University Press, 2001.
9 Salek S. Compendium of Quality of Life Instruments. Chichester: John Wiley & Sons, 1998.
10 Patient Reported Outcome and Quality of Life Instruments Database. Available from: http://www.qolid.org [Accessed August 21, 2006].
11 Biblioteca virtual de Instrumentos de Resultados Percibidos de Red IRYSS (Investigación sobre Resultados de Salud y Servicios Sanitarios). Available from: http://www.rediryss.net [Accessed August 21, 2006].
12 Badia X, Salamero M, Alonso J. La Medición de la Salud. Guía de escalas de medición en español. Barcelona: Edimac, 2002.
13 Badia X, Baro E. Cuestionarios de salud en España y su uso en atención primaria. Aten Primaria 2001;28:349–56.

14 Lohr KN, Aaronson NK, Alonso J, et al. Evaluating quality-of-life and health status instruments: development of scientific review criteria. Clin Ther 1996;18:979–92.
15 Scientific Advisory Committee of the Medical Outcomes Trust. Assessing health status and quality-of-life instruments: attributes and review criteria. Qual Life Res 2002;11:193–205.
16 Valderas JM, Ferrer M, Alonso J. [A checklist for health related quality of life measures and other patient reported outcomes.] Med Clin (Barc) 2005;125(Suppl. 1):S56–60.
17 Nunnally JC, Bernstein IH. Psychopsychometric Theory (3rd ed.). New York: McGraw-Hill, 1994.
18 Streiner DL, Norman GR. Health Measurement Scales (2nd ed.). Oxford: Oxford University Press, 1995.
19 The AGREE Collaboration. Development and validation of an international appraisal instrument for assessing the quality of clinical practice guidelines: the AGREE project. Qual Saf Health Care 2003;12:18–23.
20 Burgers JS, Fervers B, Haugh M, et al. International assessment of the quality of clinical practice guidelines in oncology using the Appraisal of Guidelines and Research and Evaluation Instrument. J Clin Oncol 2004;22:2000–7.
21 Boluyt N, Lincke CR, Offringa M. Quality of evidence-based pediatric guidelines. Pediatrics 2005; 115:1378–91.
22 Gaebel W, Weinmann S, Sartorius N, et al. Schizophrenia practice guidelines: international survey and comparison. Br J Psychiatry 2005;187:248–55.
23 Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992;30:473–83.
24 Alonso J, Prieto L, Antón JM. La version española del SF-36 Health Survey (Cuestionario de Salud SF-36): un instrumento para la medida de resultados clínicos. Med Clin (Barc) 1995;104:771–6.
25 Hunt SM, McEwen J. The development to a subjective health indicator. Sociol Health Illn 1980;2:231–46.
26 Alonso J, Anto JM, Moreno C. Spanish version of the Nottingham Health Profile: translation and preliminary validity. Am J Public Health 1990;80:704–8.
27 Nelson E, Wasson J, Kirk J, et al. Assessment of function in routine clinical practice: description of the COOP-Chart method and preliminary findings. J Chronic Dis 1987;40(Suppl. 1):55S–63S.
28 Lizán L, Reig FA. Adaptación transcultural de una medida de la calidad de vida relacionada con la salud: la versión española de las viñetas COOP/WONCA. Aten Primaria 1999;24:75–82.
29 EuroQol Group. EuroQol—a new facility for the measurement of health-related quality of life. Health Policy 1990;16:199–208.
30 Badia X, Fernández E, Segura A. Influence of sociodemographic and health status variables on evaluation of health states in a Spanish population. Eur J Public Health 1995;5:87–93.
31 Aaronson NK, Ahmedzai S, Bergman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst 1993;85:365–76.
32 Arraras JI, Illarramendi JJ, Valerdi JJ. El cuestionario de calidad de vida para cáncer de la EORTC, QLQ-C30. Estudio estadístico de validación con una muestra española. Revista Psicología la Salud 1995;7:13–33.
33 The Cochrane Collaboration. The Cochrane Manual Issue 2, 2005. Available from: http://www.cochrane.org/admin/manual.htm [Accessed March 12, 2005].
34 Pubmed. Available from: http://www.pubmed.gov [Accessed August 21, 2006].
35 Valderas JM, Mendivil J, Parada A, et al. [Development of a geographic filter for PubMed to identify studies performed in Spain]. Rev Esp Cardiol 2006;59:1244–51.
36 ISI Web of Knowledge. Available from: http://portal.isiknowledge.com [Accessed August 21, 2006].
37 Emery MP, Perrier LL, Acquadro C. Patient-reported outcome and quality of life instruments database (PROQOLID): frequently asked questions. Health Qual Life Outcomes 2005;3:12.
38 Latin American and Caribbean Health Sciences Literature (LILACS). Available from: http://www.bireme.br/abd/I/homepage.htm [Accessed August 21, 2006].
39 Scientific Library Online (SciELO). Available from: http://www.scielo.org [Accessed August 21, 2006].
40 Índice Médico Español (IME). Available from: http://bddoc.csic.es:8080/IME/BASIS/ime/web/docu/SF [Accessed January 10, 2007].
41 Base de Datos Teseo. Consejo de Coordinación Universitaria. Available from: http://www.mcu.es/TESEO/teseo.html [Accessed August 21, 2006].
42 Tesis Doctorales en Red (TDR). Available from: http://www.tdr.cesca.es [Accessed August 21, 2006].
43 BiblioPRO. Biblioteca virtual de Instrumentos de Medida de Resultados Percibidos por los Pacientes. Available from: http://iryss.imim.es/iryss/BiblioPRO.asp [Accessed August 21, 2006].
44 Banco de Instrumentos Psicométricos BIPFAES. Available from: http://bipfaes.faes.es [Accessed August 21, 2006].
45 McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? Qual Life Res 1995;4:293–307.
46 Fayers P, Hays R, eds. Assessing Quality of Life in Clinical Trials (2nd ed.). Oxford: Oxford University Press, 2005.
47 Prieto L, Lamarca R, Casado A. La evaluación de la fiabilidad en las observaciones clínicas. Med Clin (Barc) 1998;110:142–5.
48 ISI Journal Citation Reports. 2003 JCR Edition. Available from: http://www.accesowok.fecyt.es/jcr/ [Accessed January 7, 2007].
49 Armitage P, Berry G, Matthews JNS. Statistical Methods in Medical Research. Boston: Blackwell Publishing, 2002.
50 R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing, 2005. ISBN

3-900051-07-0. Available from: http://www.R-project.org [Accessed December 27, 2006].
51 Hays RD, Brown J, Brown LU, et al. Classical test theory and item response theory analyses of multi-item scales assessing parents' perceptions of theirchildren's dental care. Med Care 2006;44(Suppl. 3):S60–8.
52 Terwee CB, Bot SD, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007;60:34–42.
53 Valderas JM, Kotzeva A, Espallargues M, et al. The impact of measuring patient-reported outcomes in clinical practice: a systematic review of the literature. Qual Life Res 2008: in press.
54 Food and Drug Administration. Draft guidance for industry on patient-reported outcome measures: use in medicinal product development to support labeling claims. Fed Regist 2006;71:5862–3.