

Instrumentos de medida de calidad de vida relacionada con la salud y de otros resultados percibidos por los pacientes

José María Valderas^a, Montse Ferrer^a y Jordi Alonso^{a,b}

^aUnidad de Investigación en Servicios Sanitarios. Institut Municipal d'Investigació Mèdica (IMIM-IMAS). Barcelona. España.

^bUniversitat Autònoma de Barcelona (UAB). Barcelona. España.

En los últimos años se ha consolidado la investigación en resultados percibidos por los pacientes en general, y de calidad de vida relacionada con la salud en particular, con el desarrollo de un número elevado de instrumentos de medida. Siguiendo las recomendaciones del comité científico del Medical Outcomes Trust, el listado de comprobación ePRO propone una evaluación modular de estos instrumentos basada en 8 atributos básicos: el modelo conceptual y de medida, la fiabilidad, la validez, la sensibilidad al cambio, la interpretabilidad de las puntuaciones, la carga de administración para el entrevistador y el entrevistado, los formatos alternativos al estándar y el proceso de adaptación transcultural. Para cada atributo se lista un número variable de criterios que se debería considerar en su desarrollo. Es de esperar que la adopción de guías comunes facilite el desarrollo de mejores instrumentos de medida de resultados percibidos por los pacientes.

Palabras clave: Calidad de vida relacionada con la salud. Resultados percibidos por los pacientes. Evaluación. Validez. Fiabilidad. Cuestionarios.

Health-related quality of life instruments and other patient-reported outcomes

In the last few years, many different instruments assessing health-related quality of life and the broader concept of Patient-Reported Outcomes (PRO) have been developed. Following the recommendations of the Scientific Advisory Committee of the Medical Outcomes Trust, the ePRO checklist provides a modular assessment for PRO instruments based on eight key attributes: conceptual and measurement model, reliability, validity, responsiveness, interpretability, respondent and administrative burden, alternative forms, and cross-cultural adaptation. Selected criteria describing each attribute in detail are also listed. The availability of explicit guidelines is expected to promote the development of better PRO measurement instruments.

Key words: Health-Related Quality of Life. Patient Related Outcomes. Evaluation. Validity. Reliability. Questionnaires.

Introducción

En los últimos 30 años, se ha consolidado el área de investigación en los resultados percibidos por los pacientes, especialmente en el ámbito de la medida de la Calidad de Vida Relacionada con la Salud (CVRS)¹. Gran parte de los esfuerzos se ha consagrado a la construcción de instrumentos de medida con propósitos diversos. Los más de 1.270 instrumentos de resultados percibidos por los pacientes identificados en el ámbito internacional y los más de 350 concebidos para su uso en población española varían notablemente

Trabajo financiado parcialmente por el Instituto de Salud Carlos III (Red IRYSS G03/202 y Exp N° 3000118) y la Generalitat de Catalunya 2001 (SGR0405).

Correspondencia: Dr. J. Alonso.
Unidad de Investigación en Servicios Sanitarios.
Institut Municipal d'Investigació Mèdica (IMIM-IMAS).
Doctor Aiguader, 80. 08003 Barcelona. España.
Correo electrónico: jalonso@imim.es

en cuanto a su contenido, los métodos seguidos para su obtención, sus propiedades métricas, las necesidades para su aplicación y su uso efectivo^{2,3}.

Se han realizado numerosas propuestas para sistematizar la evaluación de estos instrumentos tanto por parte de investigadores independientes⁴⁻⁸ como auspiciadas por sociedades científicas (Internacional Society for Quality of Life Research⁹), agencias gubernamentales (Agencia Europea para la Evaluación de Medicamentos¹⁰), una acción coordinada (proyecto ERIQA, European Regulatory Issues on Quality of Life Assessment Group¹¹) o instituciones privadas (Medical Outcomes Trust^{12,13}).

En el ámbito español, destaca la Red de Investigación en Resultados y Servicios Sanitarios (Red IRYSS)³, una de cuyas líneas de investigación incluye la evaluación sistemática de los instrumentos de medida de CVRS disponibles para su uso en España. Los autores de este artículo han tenido una participación muy activa en esta línea que se concreta en el proyecto ePRO (acrónimo en inglés, que corresponde a Evaluación de Resultados Percibidos por los Pacientes). El proyecto ePRO ha consistido en el desarrollo y aplicación de un procedimiento estandarizado de evaluación de instrumentos de medida.

Selección de la fuente principal de la lista de comprobación

Los criterios usados en el proyecto ePRO se basan casi totalmente en la propuesta del comité científico del Medical Outcomes Trust¹³. Esa propuesta ha sintetizado las propuestas previas y ha sistematizado los atributos y criterios pertinentes para la evaluación de un instrumento de CVRS. Esta síntesis destaca por su actualidad, el carácter sistemático y el suficiente detalle con que explicita cada criterio. Es por ello por lo que también en el presente artículo se propone seguir dicha guía.

La propuesta diferencia 8 atributos fundamentales, que incluyen el modelo conceptual y de medida, las características psicométricas clásicas (fiabilidad, validez y sensibilidad al cambio), la interpretabilidad de las puntuaciones, la carga que la administración del instrumento implica para el entrevistador y el entrevistado y la disponibilidad y las características de formatos alternativos al estándar. También se incluye el proceso de adaptación transcultural que se ha seguido, cuando el original se ha desarrollado en un medio distinto. A cada atributo le corresponde una serie de criterios que han de permitir su valoración precisa.

El procedimiento de construcción de un instrumento de medida de resultados percibidos por el paciente es complejo y de naturaleza iterativa. Ello hace que normalmente los atributos de un instrumento se evalúen en distintos artículos científicos. Por ello, el listado de comprobación se ha construido de forma modular, para que se pueda aplicar únicamente a los atributos que específicamente se pretenda eva-

luar en un momento dado. A continuación se define y comenta con mayor detalle cada uno de estos atributos que, junto con los criterios de evaluación, aparecen en la tabla 1.

Atributos de los instrumentos de resultados percibidos por los pacientes y criterios para su evaluación

Modelo conceptual y de medida

El modelo conceptual que subyace a la construcción y uso de un instrumento se puede entender como la descripción razonada de los conceptos y poblaciones que se pretende evaluar, así como de las relaciones entre esos conceptos. Este modelo conceptual, teórico, se materializa en la práctica mediante el modelo de medida, en el que se reflejan en la estructura general del instrumento y de cada escala los procedimientos que se siguen para la obtención de las puntuaciones globales y por escalas.

En relación con el modelo de medida, es necesario que cada escala mida un único constructo (unidimensionalidad), que las diversas escalas midan constructos distintos y que la escala represente adecuadamente la variabilidad observable. Para la evaluación de la dimensionalidad de las escalas y el modelo empírico de medida se puede utilizar aproximaciones basadas en la teoría clásica de los tests, como el análisis factorial y de componentes principales¹⁴. De forma alternativa, se puede aplicar métodos basados en la Teoría de Respuesta al Ítem (TRI) o en modelos de ecuaciones estructurales¹⁵⁻¹⁷.

Para la evaluación de la variabilidad, se suele emplear como indicadores el efecto techo (proporción de individuos que alcanzan la puntuación máxima) y el efecto suelo (proporción de individuos con la puntuación mínima); en ambos casos se recomiendan valores inferiores al 15%¹⁸. Finalmente, en ocasiones, los investigadores han realizado laboriosos ejercicios de ponderación para establecer el valor relativo de cada ítem o escala, pero lo más frecuente es que no se aplique ningún tipo de pesos, y se presume entonces de forma implícita su equivalencia⁸.

Fiabilidad

La fiabilidad se ha definido tradicionalmente como el grado en que un instrumento está libre de error aleatorio, una conceptualización complementaria a la que en epidemiología relaciona la validez con la ausencia de error sistemático. Las fuentes de error aleatorio se han buscado en las respuestas proporcionadas a los diferentes ítems que componen una misma escala en una misma administración del instrumento (consistencia interna), entre las distintas administraciones de un mismo instrumento a la misma población (fiabilidad test-retest) o entre distintos observadores o administradores (fiabilidad interobservador)^{5,19}.

Para la evaluación de la consistencia interna, el grado en que todos los ítems de una escala miden el mismo constructo, se ha empleado habitualmente el coeficiente alfa de Cronbach y, con menor frecuencia, el coeficiente de Kuder Richardson KR-20, para opciones de respuesta dicotómicas. Se han propuesto como estándares mínimos apropiados valores de alfa de Cronbach de 0,70 para realizar comparaciones entre grupos. Para comparaciones individuales, el estándar mínimo se fija en 0,90 e incluso 0,95, porque de otra forma el intervalo de confianza de las puntuaciones individuales resulta tan amplio que el uso del instrumento resultaría ineficaz¹⁹. En la aproximación según teoría de respuesta al ítem, la fiabilidad se mide para cada puntuación, pero es posible calcular un parámetro, la fiabilidad marginal, de significado análogo al coeficiente de fiabilidad clásico¹³.

Los otros dos tipos de fiabilidad se conocen también de for-

TABLA 1

Listado de comprobación ePRO: atributos y criterios para la evaluación de instrumentos de resultados percibidos por los pacientes

Modelo conceptual y de medida	<input type="checkbox"/> Concepto que se pretende medir <input type="checkbox"/> Fundamentos para la obtención de los ítems y su combinación <input type="checkbox"/> Dimensionalidad del instrumento y distinción de escalas <input type="checkbox"/> Participación de la población diana en la obtención del contenido <input type="checkbox"/> Variabilidad de la escala en la población <input type="checkbox"/> Justificación de la escala de medida <input type="checkbox"/> Fundamento para el cálculo de las puntuaciones
Fiabilidad	
a) Consistencia interna	<input type="checkbox"/> Métodos para la estimación de la consistencia interna <input type="checkbox"/> Estimación de la consistencia interna según la <i>teoría clásica de los tests</i> , con estimadores centrales y error estándar de cada puntuación (dimensiones y/o total) <input type="checkbox"/> Estimación de la consistencia interna según la <i>teoría de respuesta</i> a los ítems, con EEM en el rango de puntuaciones y fiabilidad marginal de cada dimensión <input type="checkbox"/> Consistencia interna en cada población de interés
b) Reproducibilidad	<input type="checkbox"/> Métodos para la estimación de la reproducibilidad <input type="checkbox"/> Fundamento y diseño de la comparación test-retest, con intervalo que garantice la estabilidad de la población <input type="checkbox"/> Estimación sobre la reproducibilidad test-retest e interevaluador mediante CCI <input type="checkbox"/> Información sobre la comparabilidad de los parámetros del ítem (dificultad) y de la precisión de la medida en administraciones repetidas
Validez	<input type="checkbox"/> Fundamento para la estimación de la validez <input type="checkbox"/> Métodos para la estimación de la validez <input type="checkbox"/> Características de las muestras empleadas para la estimación de la validez <input type="checkbox"/> Estimaciones de la validez en cada población de interés <input type="checkbox"/> Formulación de hipótesis <i>a priori</i> y su evaluación (<i>validez de constructo</i>) <input type="checkbox"/> Fundamentos para la elección del criterio (<i>validez de criterio</i>)
Sensibilidad al cambio	<input type="checkbox"/> Métodos para la estimación de la sensibilidad al cambio (población, intervención, intervalo) <input type="checkbox"/> Diseño longitudinal con comparación de un grupo estable y un grupo que haya cambiado <input type="checkbox"/> Estimación de diferencias en las puntuaciones
Interpretabilidad	<input type="checkbox"/> Fundamentos para la elección de los criterios externos o poblaciones <input type="checkbox"/> Información sobre la forma de presentar los resultados <input type="checkbox"/> Datos comparativos o normativos de referencia
Carga	
a) Para el entrevistado	<input type="checkbox"/> Información sobre las necesidades de administración (tiempo, habilidades de lectura y comprensión, necesidades específicas) <input type="checkbox"/> Demostración de que la administración no supone un esfuerzo físico o emocional inadecuado <input type="checkbox"/> Información sobre las circunstancias en las que la administración no es apropiada
b) Para el entrevistador	<input type="checkbox"/> Información sobre las necesidades de administración (tiempo, habilidades y necesidades específicas de formación u otro tipo)
Formatos de administración	<input type="checkbox"/> Información sobre fiabilidad, validez, interpretabilidad y carga para cada formato de administración distinto <input type="checkbox"/> Información sobre la comparabilidad entre los distintos formatos
Adaptación cultural y lingüística (sólo para instrumentos adaptados, ver tabla 2)	

CCI: coeficiente de correlación intraclase; EEM: error estándar de la media. Modificado de: Scientific Advisory Committee of the Medical Outcomes Trust¹³.

ma general como reproducibilidad. La reproducibilidad test-re-test se evalúa mediante la administración repetida del instrumento de CVRS a una misma población que se haya mantenido estable en relación con el constructo objeto de medición. Con frecuencia, sólo es posible asegurar esta estabilidad entre administraciones inmediatas muy repetidas, por lo que en general se selecciona en un lapso que ronda las 2 semanas. En la evaluación de la reproducibilidad test-retest e interobservador se emplean medidas de la concordancia, habitualmente el coeficiente de correlación intraclass, por tratarse de medidas continuas²⁰. Los estándares mínimos recomendados son los mismos que para la evolución de la consistencia interna^{7,19}.

Validez

La definición más aceptada de validez la identifica con el grado en que un instrumento de CVRS efectivamente mide lo que se supone que mide. De una manera algo menos genérica se podría entender como el grado en que las puntuaciones obtenidas con el instrumento reflejan adecuadamente la CVRS del individuo o de las poblaciones a las que se aplica. Se consideran 3 tipos distintos de validez: *a)* contenido, *b)* constructo y *c)* criterio^{5,13}.

La validez de contenido (*a*) se refiere a la demostración de su adecuación al propósito específico del instrumento, esto es al grado en que se incorpora el fenómeno que se pretende medir. Suele evaluarse mediante la participación de paneles de expertos y de pacientes. El estudio de la validez de constructo (*b*) se basa en la descripción de las complejas interrelaciones entre las medidas realizadas con el instrumento y la constelación de conceptos integrados en el modelo conceptual. Tradicionalmente se ha diferenciado entre validez de constructo convergente (*b1*) o estudio de la correlación con medidas similares, y validez de constructo divergente, orientado a la identificación de falta de asociación con medidas dispares o de diferencias en las puntuaciones entre grupos diferenciados o extremos según criterios clínicos o epidemiológicos. Las matrices multirrasgo-multimétodo suelen ser la técnica de elección para el estudio de las asociaciones, pero para su correcta interpretación resulta imprescindible la formulación de hipótesis *a priori* sobre los patrones de relación esperados²¹. Se ha propuesto que valores de correlación mayores o iguales que 0,60 se corresponderían con una asociación de elevada intensidad y que los menores de 0,30 indicarían falta de asociación²².

La validez de criterio (*c*), finalmente, se refiere al grado en que las medidas obtenidas con el instrumento de CVRS se asemejan a las obtenidas con un método de referencia. La falta de medidas de referencia ha hecho que, hasta la fecha, la evaluación de este tipo de validez se haya aplicado casi exclusivamente al estudio de versiones reducidas de instrumentos, empleando el instrumento original como estándar.

Sensibilidad al cambio

Este concepto se refiere a la capacidad del instrumento para detectar diferencias en la magnitud del constructo. En ocasiones también se ha definido como el cociente entre la señal, entendida como diferencia real entre medidas, y el ruido, entendido como variabilidad de la medida²³.

En su evaluación se suele comparar las puntuaciones obtenidas antes y después de una intervención que se entiende que tendrá como efecto una modificación en la magnitud del constructo, cuantificada mediante estimaciones estandarizadas de las diferencias (tamaño del efecto)²³. Para ello se han propuesto diversos índices estadísticos que ponen

en relación las diferencias con una medida de variabilidad de la medida, todos ellos con fortuna variable y sin que ninguno haya resultado preferido a los demás²⁴. Otros métodos habitualmente utilizados son la comparación de las puntuaciones con una pregunta general de cambio y el anclaje de las puntuaciones, esto es su relación con cambios en otras medidas, ya sea con carácter transversal o longitudinal, o con componentes de las propias medidas (p. ej., porcentaje de pacientes capaces de realizar una actividad contenida en el propio instrumento)²⁵.

La clasificación de este concepto como característica independiente es, además, objeto de debate²⁴. Por un lado, la capacidad de discriminación de cambios en la magnitud del constructo podría acomodarse sin problemas como forma de evaluación longitudinal de la validez. Por otra parte, también está claro que la fiabilidad del instrumento tiene un importante efecto en la cantidad de «ruido». Finalmente, las estimaciones del tamaño del efecto y la relación de las puntuaciones con otras medidas están íntimamente asociadas a la interpretación de las puntuaciones, como se verá en el siguiente apartado.

Interpretabilidad

Se entiende por interpretabilidad el grado en que es posible asignar un significado comprensible a las puntuaciones de un instrumento de CVRS y que se facilita por la posibilidad de relacionar las puntuaciones (o diferencias en esas puntuaciones) con otras medidas inteligibles. Como se deducirá, el mismo uso repetido del instrumento lo hace interpretable, por cuanto permite establecer esas relaciones, ya sea explícita o implícitamente^{26,27}.

La supuesta falta de interpretabilidad de las medidas de CVRS es consecuencia directa de su novedad, más que una característica intrínseca de lo que se mide. Se han propuesto diversas maneras de incrementar las puntuaciones, como la publicación de la distribución de las puntuaciones en muestras representativas de la población general, la realización de estudios que utilicen una gran variedad de medidas de CVRS y otras relacionadas, el estudio de la asociación entre las puntuaciones (o sus diferencias) con determinados diagnósticos con la necesidad de tratamiento o con el efecto de tratamientos bien establecidos y también su asociación o su capacidad de predicción de ciertos eventos vitales, entre otros¹. Recientemente también se han formulado propuestas de interpretación según el contenido basadas en TRI, cuya relevancia reside en el hecho de que no precisen información externa al propio instrumento²⁷.

Carga para el entrevistador y para el entrevistado

La carga de administración se define como el tiempo, el esfuerzo y las necesidades y exigencias derivadas de la administración de un instrumento que recaigan sobre la persona a quien se aplica (carga para el entrevistado) o sobre los terceros que intervienen (carga para el entrevistador).

Se mide en relación con consideraciones de índole operativa, como el tiempo que se tarda en administrar el cuestionario, pero también en relación con los requisitos mínimos para una administración efectiva, como la capacidad de lectura y comprensión necesaria por parte del entrevistado. Un aspecto complementario a éstos y en el que no se suele reparar lo suficiente es la aceptabilidad por parte del entrevistado de las áreas de evaluación y de las formulaciones específicas empleadas en el instrumento, como podría ser el caso de preguntas relacionadas con la actividad sexual¹³.

Formatos de administración

Las formas básicas de administración de los instrumentos de medida de CVRS incluyen la autoadministración (supervisada o no), la administración mediante personal entrenado (telefónica o en persona), ambas a su vez asistidas o no por ordenador y la administración a personas próximas distintas del paciente (padres, cuidadores u otros). Cada una de ellas presenta ventajas e inconvenientes que se debe tener en cuenta a la hora de seleccionar un instrumento⁶. Para una descripción suficiente del instrumento es necesario señalar, además de la forma estándar, las formas alternativas de administración de que se dispone¹³.

Un reto específico para la estandarización: la comparabilidad transcultural

La adaptación cultural y lingüística es el octavo criterio propuesto por el comité científico del Medical Outcomes Trust. Este criterio es especialmente relevante para el ámbito español pues la mayoría de los instrumentos de medida de CVRS de que se dispone son adaptaciones de cuestionarios originalmente desarrollados en otros países. Además, buena parte de los artículos científicos en esta área describen ese proceso. Con el fin de garantizar la aplicabilidad de estos instrumentos a otras poblaciones en general, y a la española, en particular, es preciso alcanzar la equivalencia conceptual y lingüística con el original. Se establece la primera cuando los conceptos medidos son equivalentes en cuanto a significado y relevancia en la aplicación de los instrumentos a las poblaciones diana respectivas^{28,29}.

La equivalencia lingüística se refiere, en cambio, a la formulación de esos conceptos de forma que las estructuras semánticas seleccionadas cumplan la misma función y posean el mismo significado, tanto para cada punto como para las opciones de respuesta y todos los otros aspectos del instrumento^{30,31}. Repárese en que la equivalencia lingüística trasciende la simple traducción: un instrumento desarrollado en España en castellano no tendrá necesariamente aplicabilidad inmediata en cualquier otro país de habla hispana. Para alcanzar los tipos de equivalencia descritos conviene seguir un proceso cuidadoso de adaptación transcultural, que ha sido objeto de sistematización repetida^{8,28,29,32} (tabla 2). En

TABLA 2

Listado de comprobación ePRO de aplicación exclusiva para instrumentos adaptados

<p>Adaptación cultural y lingüística</p> <p>a) Traducción</p> <p><input type="checkbox"/> Traducción de los ítems y de las opciones de respuesta del instrumento original (O): al menos 2 versiones independientes (T1)</p> <p><input type="checkbox"/> Conciliación y síntesis de las versiones (T1) en una versión consensuada (T1_c)</p> <p>b) Traducción inversa (retrotraducción)</p> <p><input type="checkbox"/> Traducción de la versión consensuada (T1_c) al idioma original: retrotraducción (R)</p> <p><input type="checkbox"/> Conciliación de diferencias entre el instrumento original (O) y la retrotraducción (R)</p> <p>c) Versión predefinitiva</p> <p><input type="checkbox"/> Síntesis de versiones previas y evaluación de la equivalencia lingüística (semántica, idiomática) y conceptual</p> <p><input type="checkbox"/> Elaboración de la versión adaptada predefinitiva (T2)</p> <p>d) Ensayo de administración</p> <p><input type="checkbox"/> Administración de la versión adaptada predefinitiva (T2) a muestra de conveniencia</p> <p><input type="checkbox"/> Entrevista cognitiva para evaluación de la comprensión y aplicabilidad de la versión adaptada predefinitiva (T2)</p>

general se considera recomendable realizar al menos 2 traducciones independientes del instrumento del idioma original al de adaptación (preferentemente más) y al menos una traducción inversa (también llamada retrotraducción) de la anterior de nuevo al idioma original. Es muy conveniente la participación en los estudios de adaptación de población bilingüe en ambos idiomas y con un cierto grado de conocimiento de ambas culturas. Finalmente, cabe insistir en que, a los efectos de todas las características descritas a continuación, tanto psicométricas como de otra índole, se debe considerar de forma diferenciada el instrumento original y *cada una* de sus adaptaciones^{5,11,13,19}.

Los criterios descritos en la tabla 2 deberían considerarse de mínimos. De hecho, las propuestas más detalladas³² exigen evidencia empírica de que la versión es aceptable y comprensible para la nueva cultura (lo que requiere la participación de individuos de la población general y grupos de pacientes), así como un esfuerzo específico de armonización internacional (es decir, con otras versiones culturales del mismo instrumento). Obviamente estos criterios ya se tienen en cuenta en los instrumentos que se desarrollan originalmente en diversos países de manera simultánea, como por ejemplo el WHOQOL³³.

Conclusión

Este listado de comprobación se ha concebido como guía para autores en la elaboración de protocolos para la construcción de instrumentos de resultados percibidos por los pacientes, así como para orientar la presentación de los métodos empleados y de sus resultados, tanto a los autores como a los evaluadores de futuras publicaciones. Es de esperar que la adopción de unas guías comunes facilite la labor de todos los implicados y permita alcanzar el objetivo de disponer de mejores herramientas de medida de resultados percibidos por los pacientes.

REFERENCIAS BIBLIOGRÁFICAS

- 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.
- Garratt A, Schmidt L, Mackintosh A, Fitzpatrick R. Quality of life measurement: bibliographic study of patient assessed health outcome measures. *BMJ.* 2002;324:1417.
- BiblioPRO. Biblioteca virtual de Instrumentos de Resultados Percibidos de Red IRYSS (Investigación sobre Resultados de Salud y Servicios Sanitarios) [citado 30 jun de 2005]. Disponible en: <http://www.rediryss.net>
- McDowell I, Newell C. *Measuring Health. A guide to rating scales and questionnaires.* New York: Oxford University Press; 1996.
- Streiner DL, Norman GR. *Health measurement scales: a practical guide to their development and use.* 2nd ed. New York: Oxford; 1995.
- Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. *Ann Intern Med.* 1993;118:622-9.
- Tully MP, Cantrill JA. *Subjective outcome measurement: a primer.* *Pharm World Sci.* 1999;21:101-9.
- Badía X, Salamero M, Alonso J. *La medida de la salud. Guía de escala de medición en español.* Barcelona: Edimac; 2002.
- Acquadro C, Berzon R, Dubois D, Leidy NK, Marquis P, Revicki D, et al; 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(5):522-31.
- Apolone G, De Carli G, Brunetti M, Garattini S. An Evaluation of the EMEA recommendations on the use of Quality of Life measures in drug approval. *Pharmaco Economics.* 2001;19:187-95.
- Chassany O, Sagnier P, Marquis P, Fulleton S, Aaronson N. Patient Reported Outcomes and Regulatory Issues: the Example of Health-related Quality of Life – A European Guidance Document for the Improved Integration of HRQL Assessment in the Drug Regulatory Process. *Drug Information Journal.* 2002;36:209-38.
- Lohr KN, Aaronson NK, Alonso J, Burnam MA, Patrick DL, Perrin EB, et al. Evaluating quality-of-life and health status instruments: development of scientific review criteria. *Clin Ther.* 1996;18:979-92.

13. 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.
14. Nunnally JC. *Psychometric theory.* New York: McGraw Hill; 1967.
15. Wright BD, Stone MH. *Best Test Design.* Chicago: Mesa Press; 1979.
16. Wright BD, Masters GN. *Rating Scale Analysis.* Chicago: Mesa Press; 1982.
17. RB Kline. *Principles and practice of structural equation modelling.* New York: Guilford Press; 1998.
18. McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? *Qual Life Res.* 1995;4:293-307.
19. Valderas JM on behalf of the PRO Scientific Committee, IRYSS Network. Development of a standardized assessment tool for patient reported outcomes: a pilot test with the ePRO instrument. *Qual Life Res.* 2005;14(9):2134.
20. Prieto L, Lamarca R, Casado A. La evaluación de la fiabilidad en las observaciones clínicas: el coeficiente de correlación intraclase. *Med Clin (Barc).* 1998;110:142-5.
21. Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol Bull.* 1959;56:81-105.
22. Andresen EM. Criteria for assessing the tools of disability outcomes research. *Arch Phys Med Rehabil.* 2000;81 12 Suppl 2:S1-4.
23. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. *Control Clin Trials.* 1991;12:142S-58S.
24. Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PMM. On assessing responsiveness of health related quality of life instruments: guidelines for instrument evaluation. *Qual Life Res.* 2003;12:349-62.
25. Wyrwich KW, Bullinger M, Aaronson N, Hays RD, Patrick DL, Symonds T and The Clinical Significance Consensus Meeting Group. Estimating clinically significant differences in quality of life outcomes. *Qual Life Res.* 2005;14:285-95.
26. Lydick E. Approaches to the interpretation of quality-of-life scales. *Med Care.* 2000;38 Suppl 9:1180-3.
27. Valderas JM, Alonso J, Prieto L, Espallargues M, Castells X. Content-based interpretation aids for health-related quality of life measures in clinical practice. An example for the visual function index (VF-14). *Qual Life Res.* 2004;13:35-44.
28. Bullinger M, Alonso J, Apolone G, Leplège A, Sullivan M, Wood-Dauphinee S, et al. Translating health status questionnaires and evaluating their quality: the IQOLA approach. *J Clin Epidemiol.* 1998;51:913-23.
29. Beaton DE, Bombardier C, Guillemin F, Bosi-Ferraz M. Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine.* 2000;25:3186-91.
30. Herdman M, Fox-Rushby, Badia X. A model of equivalence in the cultural adaptation of HRQoL instruments: the universalist approach. *Qual Life Res.* 1998;7:323-35.
31. Guyatt GH. The philosophy of health -related quality of life translation. *Qual Life Res.* 1993;2:461-5.
32. Wild D, Grove A, Martin M, Eremenco S, McElroy S, Verjee-Lorenz A, et al; ISPOR Task Force for Translation and Cultural Adaptation. Principles of Good Practice for the Translation and Cultural Adaptation Process for Patient-Reported Outcomes (PRO) Measures: report of the ISPOR Task Force for Translation and Cultural Adaptation. *Value Health.* 2005;8(2):94-104.
33. The WHOQOL Group. The World Health Organization Quality of Life assessment (WHOQOL): position paper from the World Health Organization. *Soc Sci Med.* 1995;41:1403-9.