2007-12

The Patient-Reported Outcomes Measurement Information System (PROMIS) Seeks to Improve and Standardize Measures of Five Generic Health-Related QOL Domains

John E. Ware Jr.
University of Massachusetts Medical School

The Patient-Reported Outcomes Measurement Information System (PROMIS) Seeks to Improve and Standardize Measures of Five Generic Health-Related QOL Domains

John E. Ware, Jr., PhD*
Health Assessment Lab and QualityMetric Incorporated Lincoln, RI, USA

KEYWORDS
PATIENT REPORTED OUTCOMES MEASUREMENT INFORMATION SYSTEM (PROMIS), ITEM BANKING, COMPUTERIZED ADAPTIVE TESTING (CAT), ITEM RESPONSE THEORY (IRT)

The Patient-Reported Outcomes Measurement Information System (PROMIS), which seeks to improve and standardize tools of clinical research across multiple NIH Institutes, is a 5-year project that is part of the US National Institutes of Health (NIH) Roadmap Initiative. More information about the NIH Roadmap Initiative is available on the Internet (go to http://nihroadmap.nih.gov). Briefly, PROMIS began in late 2004, with the formation of a multicenter cooperative network of US academic institutions and the NIH. Its purpose is to improve the tools for measuring patient-reported health outcomes in clinical research. Specifically, PROMIS collaborators are constructing and making accessible item banks that are useful in measuring key health status concepts across a wide variety of chronic conditions. The concepts are five generic health-related QOL domains: emotional distress, fatigue, pain, physical functioning, and social role participation. The approach is to develop a set of publicly-available computerized adaptive tests (CAT) for use in measuring these domains for the clinical research community. The PROMIS network includes clinicians, clinical researchers, and measurement experts and is organized around six primary research site (PRS) grants – Duke University, Stanford University, Stony Brook University, University of North Carolina at Chapel Hill, University of Pittsburgh, and University of Washington – and a Statistical Coordinating Center (CORE, Evanston Northwestern Healthcare and Northwestern University). NIH scientists also participate in the PROMIS network, and a Scientific Advisory Board oversees PROMIS and evaluates its progress toward its goals.

The initial goals of PROMIS are documented in detail in recent publications[1], and its progress is updated regularly on the Internet (go to www.nihpromis.org) and in public meetings (the second public meeting is planned for March[3-5], 2008 in Bethesda, Maryland, USA). Briefly, PROMIS seeks to: (1) Develop pools of questionnaire items from legacy questionnaires and newly-developed items measuring five generic health-related QOL domains (emotional distress, fatigue, pain, physical functioning, and social role participation); (2) Collect questionnaire data from large samples of adults and children including those who are well and those in the general population with selected chronic diseases; (3) Analyze questionnaire and other data to calibrate and study items in each bank; (4) Build an electronic Web-based repository to administer and report health assessments using computerized adaptive testing (CAT) methods; (5) Conduct feasibility studies to evaluate CAT-based and "static" forms derived from the PROMIS item banks along with other research projects that will benefit PROMIS; and (6) Develop a plan to establish a public-private partnership to sustain the item bank repository, improve data collection, add domains and items, test the items in new populations, maintain the system in the public domain, and expand the applications of the system in clinical research and practice.

Development of the PROMIS item banks followed a common protocol across PRS teams, within the context of an established domain hierarchy developed by the PROMIS network[2]. The item pools include items from legacy questionnaires, (e.g., Health Assessment Questionnaire[3] and the SF-36® Health Survey[4] in the case of Physical Functioning), along with newly created items hypothesized to improve each bank. Item pools were developed first by a systematic search for existing items in currently available scales. Expert item review and revision was conducted to review the wording of each item and revise it as appropriate within PROMIS conventions, as well as to write additional items needed to expand the item pool. Focus groups and cognitive interviews were used to provide patient input into the domain definition and item wording process.

In addition to the qualitative studies referred to above, large-scale PROMIS data collection (N = 20,000+) has also been completed and initial data analyses have already yielded preliminary item calibrations that will be posted on the PROMIS website for all five item banks along with information on self-reported sociodemographic and clinical characteristics of study participants. Analyses of the PROMIS item banks are following a common protocol, including evaluation of

* Dr. Ware is a Co-PI (with JF Fries, MD at Stanford University) of one of the six PROMIS Primary Research Sites. Dr. Ware gratefully acknowledges the assistance of Barbara Gandek, MS in preparing this article, which expresses opinions that are entirely his own and do not necessarily reflect those of other PROMIS investigators or its sponsor.

data quality (e.g., logic and range checking) and descriptive item statistics (e.g., frequencies, means), tests of item response theory (IRT) model assumptions (e.g., unidimensionality, local independence, monotonicity), evaluation of model fit, and tests for differential item functioning[5,6].

There are good reasons to expect that this first wave of large-scale PROMIS data collection will increase our understanding of the strengths and weaknesses of widely-used legacy tools and will also enable noteworthy advances in the measurement of at least some key health domains. My confidence rests in part on the recently completed applications of IRT methods to a pre-PROMIS bank of 136 physical function items selected from nine widely-used questionnaires and administered (in various permutations) to nearly 18,000 US general population adults. Using methods similar to those we are using to analyze the first wave of PROMIS data and previously used to evaluate headache-related disability measures[7,8], my colleagues and I successfully cross-calibrated 70 of the 136 items from nine independently developed physical functioning questionnaires. Results will be published this year in the *Journal of Clinical Epidemiology*[9]. Study findings, which call into question at least some distinctions between so-called generic and disease-specific measures, suggest that forthcoming "static" and computerized dynamic physical functioning scales have the potential to substantially increase the range of measurement as well as the precision of estimates at specific score levels in head-to-head comparisons with currently widely-used "static" questionnaires of comparable length. The implications of these advances for clinical trials, namely both increasing the responsiveness of physical functioning measures *and* reducing respondent burden, has already been demonstrated in re-analyses of studies that administered questionnaire items from both the MHAQ and SF-36® Health Survey[10].

The inclusion and cross-calibration of questionnaire items from legacy tools within PROMIS item banks will greatly facilitate the interpretation of improved PROMIS metrics. Because IRT models do not yield scores that are directly interpretable, their interpretation must be determined from experience. The concurrent cross-calibration strategy of maintaining direct linkages between legacy and improved PROMIS score estimates will provide a basis for comparing new results with those from literally thousands of prior clinical studies and will facilitate much needed meta-analyses of those studies. As a result, the accumulation of interpretation guidelines will continue without interruption rather than starting all over again. An additional payoff from this approach will be a better understanding of why some scales hypothesized to measure the same domain sometimes yield different results in clinical studies.

Thus, the PROMIS approach of combining modern psychometric (i.e., IRT) and computerized dynamic methods to achieve PRO assessments that are both more practical and more precise is likely to be successful. Accordingly, the challenges faced by PROMIS, in my opinion, may be more political and administrative than scientific. What is the best way to create and promote measurement standards while making them readily available? To be well understood and accepted, standards must be well documented. To protect their scientific validity, standards and the trademarks used to label them must be copyrighted and registered so that all will know when the standards have and have not been properly reproduced. Much like the temperature scales constructed by Celsius and Fahrenheit were standardized and cross-calibrated hundreds of years ago with substantial advantages, the cross-calibration and standardization of legacy tools and improved PROMIS metrics today will substantially enable their proper use and meaningful interpretation going forward.

John E. Ware, Jr., QualityMetric Incorporated, 640 George Washington Highway, Lincoln, RI 02865, USA.
Tel: +1 (401) 334-8800; Fax: +1 (401) 334-8801;
E-mail: jware@qualitymetric.com

1. Cella D, Yount S, Rothrock N, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH Roadmap cooperative group during its first two years. *Medical Care* 2007;45(5 Suppl 1):S3-S11.
2. DeWalt DA, Rothrock N, Yount S, et al. Evaluation of item candidates: The PROMIS qualitative item review. *Medical Care* 2007;45(5 Suppl 1):S12-S21.
3. Fries JF, Spitz P, Kraines RG, et al. Measurement of patient outcome in arthritis. *Arthritis and Rheumatism* 1980;23:137:45.
4. Ware JE, Snow KK, Kosinski M, et al. SF-36 Health Survey Manual and Interpretation Guide. Boston, MA: The Health Institute, 1993.
5. Reeve BB, Hays RD, Bjorner JB, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care* 2007;45(5 Suppl 1):S22-S31.
6. Bjorner JB, Chang CH, Thissen D, et al. Developing tailored instruments: Item banking and computerized adaptive assessment. *Quality of Life Research* 2007;16(Suppl 1):95-108.
7. Ware JE, Bjorner JB, Kosinski M. Practical implications of item response theory and computerized adaptive testing: A brief summary of ongoing studies of widely used headache impact scales. *Medical Care* 2000;38(9 Suppl.):II73-82.
8. Bjorner JB, Kosinski M, Ware JE. Calibration of an item pool for assessing the burden of headaches: an application of item response theory to the headache impact test (HIT). *Quality of Life Research* 2003;12:913-33.
9. Rose M, Bjorner JB, Becker J, et al. Evaluation of a preliminary physical function item bank supports the expected advantages of the Patient-Reported Outcomes Measurement Information System (PROMIS). *Journal of Clinical Epidemiology,* 2007 (in press).
10. Martin M, Kosinski M, Bjorner JB, et al. Item response theory methods can improve the measurement of physical function by combining the modified health assessment questionnaire and the SF-36 physical function scale. *Quality of Life Research* 2007;16:647-60