"Validation" of Outcome Measures in Dermatology

Kate Viola[1], Tamar Nijsten[2] and Karthik Krishnamurthy[1]

Journal of Investigative Dermatology (2013) **133**, e13. doi:10.1038/jid.2013.332

Outcome measures are powerful tools in a clinician's armamentarium. These instruments capture clinical information and may supplement clinical judgment in order to optimize management approach, medical treatment, and referrals to other appropriate health-care providers. They may shed light on psychosocial issues while providing insight into gaps in understanding not previously considered by the clinician or the patient. These tools highlight variability between diseases when using the same scoring system and may influence clinical guideline recommendations. Additionally, these instruments may influence policy directed toward allocation of limited resources, playing a significant role in future strategies aimed at cost-effectiveness.

ADVANTAGES OF OUTCOME MEASURES

- Outcome measures quantify clinical disease severity and patient-reported outcomes.
- They are judged on the basis of their behavior when tested for certain properties, including structure, validity, reliability, and responsiveness.

LIMITATIONS

- No gold standard currently exists for comparing tools.

BACKGROUND

Scores, scales, profiles, and indexes are all examples of outcome measures. Outcome measures typically attempt to quantify either (i) clinical disease severity or (ii) patient-reported outcomes. Clinical disease severity–assessment tools gauge the global extent of disease, such as percentage of body surface area affected, physician global assessment, or the characteristics of isolated skin lesions. More disease-specific tools are the Psoriasis Area and Severity Index and the Scoring Atopic Dermatitis tool. Another group of tools focuses on patient-reported outcomes such as health-related quality of life (HRQoL), assessing the impact of a disease on patients' lives or evaluating treatment preference/satisfaction. These instruments may be generic, allowing comparison across diseases (e.g., SF-36), dermatology specific (e.g., Dermatology Quality of Life index or Skindex), disease specific, or concept specific (e.g., stigmatization or anxiety). Figure 1 demonstrates the relationships among clinical disease severity measures, HRQoL tools, and therapeutic intervention data.

HOW ARE OUTCOME MEASURES VALIDATED?

Analytical treatment of an instrument has acquired the name "validation," and it is the widely accepted method for evaluating the integrity of an instrument. The term "validation" is technically inappropriate because "validity" is only one of the axes or properties weighed. The evaluation of an instrument involves testing many properties, including validity, structure, reliability, and responsiveness.

The outcome measure must first be developed (usually a questionnaire) and administered. Next, the tool undergoes property testing (statistical analysis) to determine the integrity of the tool based on the answers that are generated from initial test administration (Weller *et al.*, 2012). Once this is accomplished with desirable results, the instrument can be used as a qualitative measure for use in a study and/or as a clinical decision–making tool.

DEVELOPMENT

Item generation

Items are questions or statements pertaining to the topic or disease condition for use within the instrument. Items may be generated in a variety of manners. "Exploratory interviews" by psychologists with a heterogeneous cohort of patients affected by the condition may be utilized to elicit significant and relevant issues of interest. Additional strategies involve systematic literature reviews of similar studies and consultation with colleagues and experts in the disease-specific field. This process generates a pool of issues that must be phrased and checked with the patients to determine whether the constructed items are interpretable and unambiguous. Items are formatted in question or statement form, followed by answer choices. Commonly chosen is the Likert 5-point scale, with answer choices such as "never"/"rarely"/"sometimes"/"often"/"always" or "strongly disagree"/"disagree"/"neutral"/"agree"/"strongly agree." It is also optional to limit these answers by instructing the patient to answer only the questions based on a specified time frame, such as the previous week, month, or year (Weller *et al.*, 2012).

[1]*Albert Einstein College of Medicine, Division of Dermatology, Department of Medicine, Bronx, New York, USA and* [2] *Department of Dermatology, Erasmus Medical Center, Rotterdam, The Netherlands*

Correspondence: Karthik Krishnamurthy, 1400 Pelham Parkway South, Building 1, Suite 4W-4D, Bronx, New York 10461, USA. E-mail: kkderm@gmail.com

Item reduction

During this stage, the large set of items is administered to a large number of target patients to determine which items carry the largest impact factor. One method involves determining the "frequency" of each item by asking patients whether they have experienced the problem described in the item over the specified time frame. The percentage of "yes" answers becomes the frequency. The patients are also asked to determine the "importance" on a 5-point scale. The product of the frequency and importance is defined as the impact factor. The higher the impact factor, the more relevant the item. Items with low impact factors can be reduced or eliminated from the instrument (Weller *et al*., 2012).

PROPERTY TESTING (VALIDATION)

Because no gold standard exists against which to compare a tool, tools are judged on the basis of their behavior when tested for certain properties, such as structure, validity, reliability, and responsiveness. For example, for assessing HRQoL, the SF-36 and Skindex behave "well" when property-tested (Both *et al*., 2007).

Structure

Many tools recognize that specific items affect specific aspects (or constructs) of a patient's life, namely, constructs within the physical domain versus the psychosocial domain; this can be further differentiated in subjective and objective impact (Muldoon *et al*., 1998). The structure of a questionnaire is important because it assesses whether the questions all address the same underlying construct (i.e., impact). Ultimately, for item/question scores to be grouped together and summed, the items should be unidimensional (i.e., measure the same underlying construct). In classic test theory, exploratory factor analysis is used as an objective method and assumes no *a priori* hypotheses regarding the construct on which an item should be loaded. It allows statistical analysis to group and associate items, with domains based on underlying patterns and relationships, without bias (Norris and Lecavalier, 2009; Fabrigar *et al*., 1999; Finch and West, 1997). Ideally, items load uniquely on one factor (dimension). If an item loads on multiple factors or on none of the extracted factors (so called "item complexity"), the item is best discarded because its significance cannot be directly attributed to only one dimension, making interpretation of its score ambiguous. The dimensionality of an instrument can also be tested using models based on the item response theory, of which the Rasch analysis is the most commonly used model in dermatology (Wright, 1996).

For instance, the Skindex includes constructs such as physical limitations and discomfort within the physical domain. Dimensions within the psychosocial domain include cognitive, social, and emotional disruptions, with the emotional dimension being further characterized by the constructs of depression, fear, embarrassment, and anger (Figure 2) (Chren *et al*., 1996).

Items must be loaded onto a specific domain for instrument results to be appropriately interpreted. For example, the Skindex item "My skin hurts" is loaded onto the "discomfort" category, a component of the physical domain. Conversely, "I think about my skin condition" is loaded on the "cognitive" category, a psychosocial domain. Items may be loaded onto a construct in a variety of ways, ranging from objective to subjective.

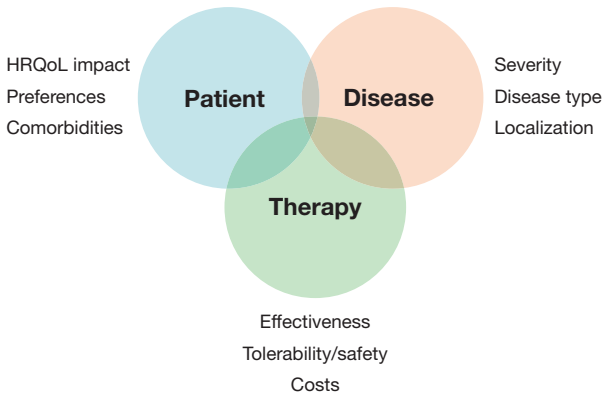

Figure 1. Clinical outcomes. Health-related quality-of-life (HRQoL) tools, the relationship among clinical disease severity–assessment measures, and therapeutic intervention data. Illustration by Tamar Nijsten.

Validity

The validity of an instrument is the extent to which it measures what it is intended to measure. Rather than a single gold standard, there are several methods with which to assess validity.

Content validity refers to the adequacy of the instrument to address all relevant items within a construct; this can be evaluated by the instrument respondents (e.g., patient). When assessed by experts, this is termed "face validity."

Convergent validity is achieved when a tool correlates well with tools that are supposed to measure the same underlying construct (e.g., a new tool assessing HRQoL in atopic eczema patients showed results similar to those obtained using an existing HRQoL tool). The statistical test used is the correlation coefficient.

Construct validity is tested by hypothesizing that different groups of patients show differences in scores as expected. For example, patients with severe disease should exhibit higher levels of HRQoL impairment than those for patients with mild disease. If this is confirmed by the outcome-measure tool, it will suggest optimal construct validity (Prinsen *et al*., 2013).

Reliability

Test–retest reliability evaluates the consistency of the score when the instrument is administered to the same person at different times, with the interval being short enough that the disease condition is unlikely to have changed. Intraclass correlation coefficients are used to determine this aspect of reliability, with 1.0 showing perfect correlation. Generally, scores greater than 0.7 are considered acceptable (Spuls *et al*., 2010).

Internal consistency can be tested after one application of an instrument and examines the degree to which the set of items measures the same construct; this is measured by Cronbach's α test. For example, an $\alpha < 0.7$ suggests that the item is not highly correlated with the other items in the

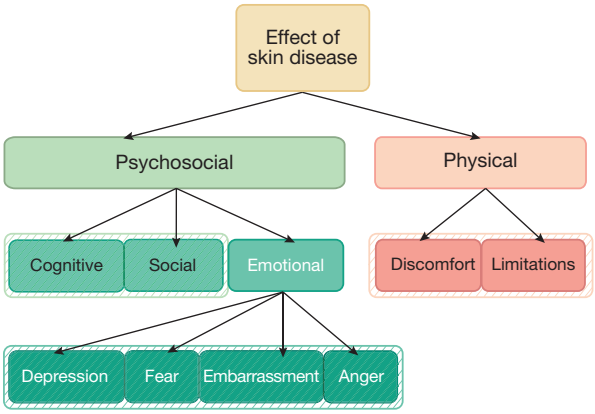

Figure 2. Conceptual framework representing the effects of skin disease on quality of life. This hypothesis was based on literature review and directed interviews with patients with skin disease and clinicians who care for them. The boxes with double borders indicate constructs addressed by the eight scales of the Skindex. Adapted from Chren *et al.* (1996).

scale, suggesting it measures a different aspect of the disease. Conversely, an $\alpha > 0.9$ suggests item redundancy, and the construct is being overemphasized and overrepresented within the instrument (Prinsen *et al.*, 2013).

Responsiveness refers to the instrument's capability to change when the patient experiences a change in disease state. Responsiveness addresses whether a tool is sensitive enough to detect changes in patients. For example, the score should be lower after a successful therapeutic intervention than it was prior to treatment. The important question is whether the change in impairment reflects a statistically significant change while actually affecting the patient. To assess this, the minimal clinical important difference can be estimated (Revicki *et al.*, 2008).

Response distribution assesses whether the entire range of the item scores is being utilized. If more than 70% of patients score an item "0" (or any other score), then this item may not discriminate between patients and may be removed from the scale.

Overall, the distribution of scores can also be measurement indicators. For example, the Psoriasis Area and Severity Index instrument curve is skewed right, underrepresenting patients who suffer from mild disease because the instrument is not as sensitive to detection of disease in this range (Spuls *et al.*, 2010).

Finally, other axes of instrument evaluation lie in identifying biases based on culture and language, as well as practical issues ranging from respondent burden (is the tool too long?) to administrative burdens, e.g., means of administration (verbal, over the phone, via computer) and data collection (Spuls *et al.*, 2010).

There are currently no guidelines for development or appropriate testing of intended health measurement within an outcome measure. The Consensus-based Standards for the Selection of Health Measurement Instruments study represents initial research in the development of a provider checklist to address this need in assessing different health-related, patient-reported outcomes using the Delphi procedure. This process includes sequential questionnaires, or "rounds," with controlled feedback to gain consensus by a group of experts. It is favorable where there is a lack of empirical evidence, yet it is able to incorporate responses from leaders in many health-care fields of expertise (Mokkink *et al.*, 2010). Thus, health-care providers may use this checklist to select appropriate measurement tools for patient feedback and optimal health-care outcomes.

WHY IS THIS IMPORTANT IN DERMATOLOGY?

With more than 50 proposed outcome measures for patients with psoriasis, how can dermatologists determine which measurement tool is appropriate for their patients (Spuls *et al.*, 2010)? The health-care provider must understand how to administer, as well as interpret the results of, an outcome measure for effective utilization while recognizing the limitations of each tool. Factors including study setting, disease manifestation, and patient type must also be considered when choosing an optimal outcome measure. Provider exposure and training are critical to a better

QUESTIONS

This article has been approved for 1 hour of Category 1 CME Credit. To take the quiz, with or without CME credit, follow the link under the "CME CREDIT" header.

1. Which of the following is NOT an example of an outcome measure?

A. Score.
B. Scale.
C. Profile.
D. Index.
E. Table.

2. Which of the following is NOT a property used to test the behavior of an outcome measurement tool?

A. Validity.
B. Structure.
C. Responsiveness.
D. Reliability.
E. Sensibility.

3. Which of the following statements regarding outcome measures is false?

A. The PASI remains the gold standard by which all other psoriasis tools are judged.
B. The COSMIN checklist is an attempt to standardize development and reporting of outcome measures.
C. Outcome measures are used to quantify clinical disease severity and patient-reported outcomes.
D. Items that load onto multiple constructs are best eliminated from the instrument.
E. In the Skindex, the emotional dimension is further subdivided into the constructs of depression, fear, embarrassment, and anger.

understanding of outcome measures and their role in determining the extent of disease burden in order to assist dermatologists in providing optimal patient care.

CONFLICT OF INTEREST

The authors state no conflict of interest.

CME CREDIT

This article has been approved for 1 hour of Category 1 CME Credit.
To take the online quiz, follow the link below:
http://www.classmarker.com/online-test/start/?quiz=yxk51dc7bff36258

SUPPLEMENTARY MATERIAL

Answers and a PowerPoint slide presentation appropriate for journal club or other teaching exercises are available at http://dx.doi.org/10.1038/jid.2013.332.

REFERENCES

Both H, Essink-Bot ML, Busschbach J *et al*. (2007) Critical review of generic and dermatology-specific health-related quality of life instruments. *J Invest Dermatol* 127:2726–39

Chren M, Lasek R, Quinn L *et al*. (1996) Skindex, a quality of life measure for patients with skin disease: reliability, validity, and responsiveness. *J Invest Dermatol* 107:707–13

Fabrigar LR, Wegener DT, MacCallum RC *et al*. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychol Methods* 4:272–99

Finch JF, West SG (1997) The investigation of personality structure: statistical models. *J Res Pers* 31:439–85

Mokkink LB, Terwee CB, Patrick DL *et al*. (2010) The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol* 63:737–45

Muldoon M, Barger S, Flory J *et al*. (1998) What are quality of life measurements measuring? *Br Med J* 316:542

Norris M, Lecavalier L (2009). Evaluating the use of exploratory factor analysis in developmental disability psychological research. *J Autism Dev Disord* 40:8–20

Prinsen CA, de Korte J, Augustin M *et al*. (2013) Measurement of health-related quality of life in dermatological research and practice: outcome of the EADV Taskforce on Quality of Life. *J Eur Acad Dermatol Venereol*, e-pub ahead of print 9 January 2013

Revicki D, Hays R, Cella D *et al*. (2008) Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol* 61:102–9

Spuls PI, Lecluse LL, Poulsen ML *et al*. (2010) How good are clinical severity and outcome measures for psoriasis?: quantitative evaluation in a systematic review. *J Invest Dermatol* 130:933–43

Weller K, Groffik A, Magerl M *et al*. (2012) Development and construct validation of the angioedema quality of life questionnaire. *Allergy* 67:1289–98

Wright BD (1996) Comparing Rasch measurement and factor analysis. *Struct Eq Model* 3:3–24