Patient-Reported Outcomes

Examining Item Content and Structure in Health Status and Health Outcomes Instruments: Toward the Development of a Grammar for Better Understanding of the Concepts Being Measured

Pennifer Erickson, PhD1,*, Richard J. Willke, PhD2
1OLGA, State College, PA, USA; 2Pfizer, Inc., New York, NY, USA

ABSTRACT

Objectives: Health outcomes instruments assess diverse health concepts. Although item-level concepts are considered fundamental elements, the field lacks structures for evaluating and organizing them for decision making. This article proposes a grammar using item stems, response options, and recall periods to systematically identify item-level concepts. The grammar uses “core concept,” “evaluative component,” and “recall period” as intuitive terms for communicating with stakeholders. Better characterization of concepts is necessary for classifying instrument content and linking it to treatment benefit.

Methods: Items in 2 generic and 21 disease-specific instruments were evaluated to develop and illustrate the use of the grammar. Concepts were assigned International Classification of Functioning, Disability and Health codes for exploring the value that the grammar and a classification system add to the understanding of content across instruments.

Results: The 23 instruments include many core concepts; emotional function is the only concept assessed in all instruments. Concepts in disease-specific instruments show obvious patterns; for example, arthritis instruments focus on physical function. The majority of instruments used the same response options across all items, with five-point scales being the most common. Most instruments used one recall period for all items. Shorter recall periods were used for conditions associated with “flares,” such as chronic obstructive pulmonary disease and “skin disease.” Every diagnosis, however, showed variation across instruments in the recall period used. Conclusions: This analysis indicates the proposed grammar’s potential for discerning the conceptual content within and between health outcomes instruments and illustrates its value for improving communication between stakeholders and for making decisions related to treatment benefit.

Keywords: Core concepts, International Classification of Functioning, Disability, and Health (ICF), patient-reported outcomes, taxonomy.

Copyright © 2013, International Society for Pharmacoeconomics and Outcomes Research (ISPOR). Published by Elsevier Inc.

Introduction

Instruments that measure health from an individual’s perspective assess a wide range of concepts of health and well-being and are used for evaluating treatment benefit throughout the world [1–4]. The development of these instruments can be traced to generic measures of health that emerged in the 1970s [5–8] and were based on the 1948 World Health Organization’s definition of health as “physical, mental, and social well-being” [9]. This tripartite classification was expanded by Ware [10] to include disease, personal functioning, psychological distress/well-being, general health perceptions, and social/role functioning. Subsequently, Wilson and Cleary [11] modified the Ware model to include a medical focus and external factors. Patrick and Chiang [12] expanded this model to include more detail on both the health concepts and factors due to the environment and personal and lifestyle factors. These conceptual models have been used as the basis for the development of numerous generic and disease-specific instruments ranging in complexity from multi-item scales that assess a single concept, for example, depression, to scales that measure multiple concepts by using multiple scales that may, or may not, be aggregated to form an overall outcome measure [13–15]. In the past decade, generic instruments have become increasingly widely used to measure health status, for example, in population surveys by the Medical Care Expenditure Survey and the Medicare Health Outcomes Survey [16,17]. During the same time, specific measures have become routinely used to assess health and treatment outcomes in clinical research studies including those used for new drug development [18].

Approximately one third of all new drugs approved by the U.S. Food and Drug Administration in the period 1997 to 2003 and one fourth in the period 2006 to 2010 included labeling claims based on at least one patient-reported outcome (PRO) instrument [18,19]. This information was used in the process of generating...
a Food and Drug Administration Guidance on PROs and a European Medicines Agency Reflection Paper [1,2]. The positions put forth in these regulatory documents are now being expanded to include information reported by clinicians and by patients’ observers. Together these reported assessments are referred to as clinical outcomes assessments (COAs) [20].

As a result of the growing importance of obtaining individuals’ input into health care decision making, the number and diversity of instruments have increased exponentially in the past 30 years. Common features, however, exist across instruments. At the item level, COAs include core concepts that are considered to be fundamental elements of the broader concepts that may be represented by multi-item scales.

Together with the core concept, other aspects of item wording are also important to fully understand item content, namely, the recall period, the response option, and the verb-adverb of the question (“how often are,” “how troubled by,” “how severe are”); these elements, which comprise the “grammar” of COAs, convey the precise way that the core concept can be systematically evaluated within a given item. Better characterization of the choice set, not only for the core concepts but also for the other aspects of the grammar, would lead to a more complete classification of the conceptual content of an item in a COA and how it measures a specific treatment benefit.

Two recent initiatives that aim to provide a structure for classifying core concepts within COAs are the Patient Reported Outcomes Measurement Information System and the International Classification of Functioning, Disability, and Health (ICF) [21,22]. Both of these add lower-level concepts, for example, activities of daily living and instrumental activities of daily living, to the physical concept. Because both the Patient Reported Outcomes Measurement Information System and the ICF were developed for uses broader than determining treatment efficacy, each lacks a structure that relates a health concept to its role in evaluating treatment outcome, for example, whether or not a concept is appropriate for making an explicit statement of treatment benefit as might appear in a label claim.

To aid the understanding of COA concepts and their contribution to the drug-approval process and decision making more broadly, Erickson et al. [23] proposed the PRO Concept Taxonomy. This hierarchical structure has four levels: family, compound, singular, and low-level singular concepts. While concepts as well as their names and definitions have yet to be specified, family-level concepts can be thought of as corresponding to those in the 1948 World Health Organization definition and its subsequent modifications. Concepts at each level correspond to the importance placed on them by those who are directly involved with the patient’s response to treatment. Some item-level concepts, such as shampooing hair, are unlikely to be of sufficient importance for labeling and would be considered low-level singular concepts. Other item-level concepts, such as pain severity, may be singular concepts if they can be used for claiming treatment benefit. Multi-item scales form singular or higher-level concepts. A hierarchical structure, such as that in the PRO Concept Taxonomy, allows investigators to identify patient-relevant concepts that are appropriate for the intended claim of treatment benefit at the time of trial design, thus increasing the likelihood that the trial results will support the desired labeling.

While other taxonomies have been proposed for use with PRO measures [24–26], these efforts focus on practical characteristics of instruments, for example, respondent burden. These systems are similar to the PRO Instrument Hierarchy that was developed as a companion structure to the PRO Concept Taxonomy. Structures that categorize instruments on the basis of their measurement features, rather than conceptual content, are also important tools that can improve communication among researchers, sponsors, and other consumers of COA information.

There is certainly an awareness of the importance of not only the core concepts but also other aspects of the grammar of COA items [27,28]. The final FDA PRO guidance devotes specific attention to recall periods and response options. There has been less effort to classify these aspects, however, than there has been to organize the concept taxonomies. Nor has there been much effort to analyze the impact of differences in recall and response formats on the measurement and interpretation of otherwise similar core concepts. Whether or not variations in these other components of the grammar will make significant differences in the overall content and psychometrics of an item is an empirical, case-by-case, question, but it cannot be assumed that they will not matter.

Better delineation of the structure and usage of this grammar would be a useful step in addressing these considerations. As this brief review indicates, the key for having meaningful discussions about the use and interpretation of COAs is to have a common language for understanding an instrument’s content and potential for decision making. This language needs to be compatible with concepts measured by existing measures and also with those in instruments being developed now and in the future.

This article evaluates the syntax and language of items to develop a descriptive, item-level, grammar of COAs that can guide users in identifying item-level concepts and understanding each concept’s role in evaluating and making a statement of treatment benefit. For comparison, analysis, and communication, core concepts are best understood as part of cohesive concept taxonomy. In addition to introducing the grammar, this article suggests the value of using both the grammar and taxonomy for evaluating instrument selection within the context of a given application. Similarities of concepts within levels of the taxonomy indicate confirmatory information; differences suggest disparate concepts that need further investigation. The grammar is designed to be applicable to existing and new and generic and disease-specific instruments. Items in a sample of PRO instruments are compared and contrasted by using the grammar to identify similarities and differences in the core concepts across instruments. The ICF was used to explore the value that the proposed grammar adds to the evaluation and classification of the conceptual content of various instruments. The ICF was selected because it has 1) a well-defined hierarchical structure similar to that in the PRO Concept Taxonomy; 2) clearly stated concept definitions that enable the matching of core concepts in items in the sampled instruments to levels in the hierarchy; and 3) a coding system that indicates the location of the concept in the taxonomy, in terms of its conceptual content and its potential role in decision making. The longer the ICF code, the more detailed the concept, the narrower the concept, and the more restricted the concept’s role in decision making. Thus, a subtext is to evaluate the ICF’s potential for populating a PRO Concept Taxonomy as a step toward developing a standardized terminology for improving communication among the various stakeholders involved in evaluating treatment benefit.

**Methods**

To systematically evaluate the conceptual content of items in COA instruments, as well as to provide considerations for new instrument development, we propose a grammar based on the item—stem wording, recall period, and response option. This grammar takes into account the role that modifying phrases play in understanding the item’s core concept. Table 1 identifies the primary components in this descriptive tool and gives a brief definition and statement of purpose for each. In addition, grammar components are expressed in terms chosen to make them more intuitive to those unfamiliar with terms used primarily by COA researchers.

All items have three components—core concept, evaluation, and recall period—with the first two each having a subcomponent.
Each item stem contains a core concept that fits within one, and only one, family-level concept and contains the facet of health, function, or well-being that is evaluated by or for the patient. Because this is usually phrased as a statement about what is happening to the patient, we refer to this as an event. An event may be conveyed by using a linking or sensory verb, for example, am, are, have, feel, or seem, or as an action, for example, drive, walk, or run. It may include an object as well as the verb, for example, “changing basic body position.” A core concept may also have a modifying phrase—a modifier—that relates the event to the disease context or to a specific causal factor, or imposes a condition on the event context; the latter does not necessarily apply attribution. This modifying phrase may be seen in the item stem itself but may also be in the instructions to the respondent. By placing the core concept/event within a specific context, the modifier allows for a more targeted, and more clearly interpreted, response. However, it also means that items with the same core concept may elicit different responses in different contexts. For example, a core concept such as being “scared or worried” could elicit quite different responses depending on whether the context is about your lung problems as compared to “about having family support when I need it.”

Each item also has an evaluation component that expresses the aspect and rating of the event. This component may appear in either the stem or the response options or both, and indicates the aspect of the core concept that is being measured, for example, frequency (how often ...?), intensity, severity, bother, and so on, as well as the scale on which it is being rated. The term aspect is used here as an expression for the nature of the impact that the event, as modified, has on the patient or respondent’s life. The specification of aspect is important, in that the interpretation of a core concept may vary significantly depending on the aspect being evaluated, such as frequency versus severity of breathlessness [29,30]. Similarly, in the rating part of this component, the wording and range of the response options may have important effects on the sensitivity of a given item, including floor and ceiling effects. Because response ranges can vary from yes-no options (e.g., the St. George’s Respiratory Questionnaire for chronic obstructive pulmonary disease [COPD]) to the visual analogue scale or 11-point ranges (e.g., the OAKHQOL for arthritis), there is a considerable margin for variation here [31,32].

The recall period, already a familiar item component, indicates the time period of the event that is relevant for the purpose of measurement, for example, determining treatment benefit or evaluating population health [33]. Recall period is sometimes expressed as an implicit “now” or as “at this point in time.”

Table 2 shows the 2 generic and 21 disease-specific instruments selected for examining the grammar components. The disease areas were chosen to reflect both common and less common disease areas; the specific instruments were chosen as being commonly used, or otherwise significant, instruments in those areas. Item stems, response options, and recall periods for these instruments were entered into a spreadsheet and evaluated to determine the core concept, including the event, modifier, evaluation aspect and rating, and recall period. For some instruments, we used information that was stated in the instructions to the respondent, when such information was not stated in the instructions. Item stems, response options, and recall periods for these instruments were entered into a spreadsheet and evaluated to determine the core concept, including the event, modifier, evaluation aspect and rating, and recall period. For some instruments, we used information that was stated in the instructions to the respondent, when such information was not stated in the individual items. This resulted in 618 items available for analysis. Table 3 presents examples of core concepts from these PRO instruments. In Chronic Respiratory Questionnaire item 12, the core concept is considered to be control or mastery; the phrase “of your breathing problems” acts as a modifier. The event of breathlessness assessed in London Chest Activities of Daily Living Scale item 1 is considered to be dependent on performing the self-care activity of drying oneself, a modifier. This item
The predominant number being compared across instruments. When there was more than one recall period, the predominant period was used. The recall periods for all items were grouped into five categories (now, last few days, past week, past month, and longer ago) and represented with four-digit codes and can be considered to correspond to low-level singular concepts in the PRO Concept Taxonomy. Shorter codes represent increasingly more composite concepts, analogous to the singular and compound concepts. In general, the ICF is quite detailed for physical functioning but relatively weak on emotional content.

Each category is accompanied by a brief description of the concept as illustrated at the bottom of Table 4. These descriptions were used for assigning ICF codes to the core concepts in our sample. Some items were assigned two codes, usually one from Functioning and Disability and the other from the Environmental Factors section, if no single code seemed to completely represent the core concept.

For this analysis, the percentage of each instrument’s items at each two-digit ICF level (e.g., d4 in Table 4) was calculated, along with the average and high and low percentages across the five instruments within a disease area. Recall periods for all items were grouped into five categories (now, last few days, past week, past 2 weeks, past month). When an instrument included more than one recall period, the predominant period was used. The number of response options for each item was identified, with the predominant number being compared across instruments.

The predominant number being compared across instruments. When there was more than one recall period, the predominant period was used. The recall periods for all items were grouped into five categories (now, last few days, past week, past month, and longer ago) and represented with four-digit codes and can be considered to correspond to low-level singular concepts in the PRO Concept Taxonomy. Shorter codes represent increasingly more composite concepts, analogous to the singular and compound concepts. In general, the ICF is quite detailed for physical functioning but relatively weak on emotional content.

Each category is accompanied by a brief description of the concept as illustrated at the bottom of Table 4. These descriptions were used for assigning ICF codes to the core concepts in our sample. Some items were assigned two codes, usually one from Functioning and Disability and the other from the Environmental Factors section, if no single code seemed to completely represent the core concept.

For this analysis, the percentage of each instrument’s items at each two-digit ICF level (e.g., d4 in Table 4) was calculated, along with the average and high and low percentages across the five instruments within a disease area. Recall periods for all items were grouped into five categories (now, last few days, past week, past 2 weeks, past month). When an instrument included more than one recall period, the predominant period was used. The number of response options for each item was identified, with the predominant number being compared across instruments.

The predominant number being compared across instruments. When there was more than one recall period, the predominant period was used. The recall periods for all items were grouped into five categories (now, last few days, past week, past month, and longer ago) and represented with four-digit codes and can be considered to correspond to low-level singular concepts in the PRO Concept Taxonomy. Shorter codes represent increasingly more composite concepts, analogous to the singular and compound concepts. In general, the ICF is quite detailed for physical functioning but relatively weak on emotional content.

Each category is accompanied by a brief description of the concept as illustrated at the bottom of Table 4. These descriptions were used for assigning ICF codes to the core concepts in our sample. Some items were assigned two codes, usually one from Functioning and Disability and the other from the Environmental Factors section, if no single code seemed to completely represent the core concept.

For this analysis, the percentage of each instrument’s items at each two-digit ICF level (e.g., d4 in Table 4) was calculated, along with the average and high and low percentages across the five instruments within a disease area. Recall periods for all items were grouped into five categories (now, last few days, past week, past 2 weeks, past month). When an instrument included more than one recall period, the predominant period was used. The number of response options for each item was identified, with the predominant number being compared across instruments.

The predominant number being compared across instruments. When there was more than one recall period, the predominant period was used. The recall periods for all items were grouped into five categories (now, last few days, past week, past month, and longer ago) and represented with four-digit codes and can be considered to correspond to low-level singular concepts in the PRO Concept Taxonomy. Shorter codes represent increasingly more composite concepts, analogous to the singular and compound concepts. In general, the ICF is quite detailed for physical functioning but relatively weak on emotional content.

Each category is accompanied by a brief description of the concept as illustrated at the bottom of Table 4. These descriptions were used for assigning ICF codes to the core concepts in our sample. Some items were assigned two codes, usually one from Functioning and Disability and the other from the Environmental Factors section, if no single code seemed to completely represent the core concept.

For this analysis, the percentage of each instrument’s items at each two-digit ICF level (e.g., d4 in Table 4) was calculated, along with the average and high and low percentages across the five instruments within a disease area. Recall periods for all items were grouped into five categories (now, last few days, past week, past 2 weeks, past month). When an instrument included more than one recall period, the predominant period was used. The number of response options for each item was identified, with the predominant number being compared across instruments.

The predominant number being compared across instruments. When there was more than one recall period, the predominant period was used. The recall periods for all items were grouped into five categories (now, last few days, past week, past month, and longer ago) and represented with four-digit codes and can be considered to correspond to low-level singular concepts in the PRO Concept Taxonomy. Shorter codes represent increasingly more composite concepts, analogous to the singular and compound concepts. In general, the ICF is quite detailed for physical functioning but relatively weak on emotional content.

Each category is accompanied by a brief description of the concept as illustrated at the bottom of Table 4. These descriptions were used for assigning ICF codes to the core concepts in our sample. Some items were assigned two codes, usually one from Functioning and Disability and the other from the Environmental Factors section, if no single code seemed to completely represent the core concept.

For this analysis, the percentage of each instrument’s items at each two-digit ICF level (e.g., d4 in Table 4) was calculated, along with the average and high and low percentages across the five instruments within a disease area. Recall periods for all items were grouped into five categories (now, last few days, past week, past 2 weeks, past month). When an instrument included more than one recall period, the predominant period was used. The number of response options for each item was identified, with the predominant number being compared across instruments.

The evaluation-aspect of the event being queried is coded for each item; common responses were frequency, severity, difficulty, and extent; more than 20 different response types were identified. The percentage of each instrument’s items with each evaluation-aspect was calculated.

---

### Table 2 – Disease or generic content and initials for PRO instruments used in this analysis.

<table>
<thead>
<tr>
<th>Context</th>
<th>Initials of PRO instruments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arthritis</td>
<td>AIMS2, DASH, HAQ, KOOS, OAKHQQL</td>
</tr>
<tr>
<td>COPD</td>
<td>CRQ, EXACT, LCADL, PFSS, SGRQ</td>
</tr>
<tr>
<td>Fibromyalgia</td>
<td>FIQ</td>
</tr>
<tr>
<td>Head &amp; neck cancer</td>
<td>FACT-H&amp;N, HNQQL, QOL-RTI, VHL, XQ</td>
</tr>
<tr>
<td>Skin disease</td>
<td>DLQI-10, PDI-15, PLSI, PSORIQol, Skinindex-29</td>
</tr>
<tr>
<td>Generic</td>
<td>EQ-5D, SF-12</td>
</tr>
</tbody>
</table>

AIMS2, Arthritis Impact Measurement Scales 2 [34]; CRQ, Chronic Respiratory Questionnaire [35]; DASH, Disabilities of the Arm, Shoulder, and Hand [36]; DLQI-10, Dermatology Life Quality Index 10 [37]; EQ-5D, EuroQol five-dimensional [38]; EXACT, EXAcerbations of Chronic Obstructive Pulmonary Disease [39]; FACT-H&N, Functional Assessment of Cancer for Patients with Head and Neck Cancer [40]; FIQ, Fibromyalgia Impact Questionnaire [41]; HAQ, Health Assessment Questionnaire [42]; HNQQL, Head and Neck Quality of Life Instrument [43]; KOOS, Knee Osteoarthritis Outcome Scale [44]; LCADL, London Chest Activities of Daily Living Scale [45]; OAKHQQL, Osteoarthritis Knee and Hip Quality of Life [32]; PDI, Psoriasis Disability Index 15 [46]; PFSS, Pulmonary Functional Status Scale [47]; PLSI, Psoriasis Life Stress Inventory [48]; PRO, patient-reported outcome; PSORIQol, Psoriasis Index of Quality of Life [49]; QOL-RTI, Quality of Life-Radiation Therapy Instrument [50]; SF-12, SF-12 Health Survey [51]; SGRQ, St. George’s Respiratory Questionnaire [31]; Skindex-29, Skindex-29 [52]; VHL, Voice Handicap Index [53]; XQ, Xerostoma Questionnaire [54].
Results

As shown in Table 5, generic and disease-specific instruments included in this analysis include a range of core concepts. The Emotional Functions category, which is in 18 of the 23 instruments, is included in each of the disease-specific categories as well as in generic instruments. Disease-specific instruments show some obvious patterns; for example, concepts of voice, communication, and swallowing are assessed only in the head and neck cancer instruments; similarly, the core concept of skin sensitivity appears only in instruments specific to skin diseases. As might be expected, the core concepts in the arthritis instruments tend to cluster in the Activity and Participation section of the ICF. The use of the Activity and Participation component, as well as mental functions (b1), however, varies quite a bit within other disease areas and is a point for consideration in comparing and evaluating instruments in those areas.

Figure 1A shows the percentage of core concepts included in the five arthritis-specific instruments according to the average-high-low analysis. Mobility was most common, comprising an average of 39% of the items across the five instruments, and was included in all five instruments here. Self-care activity was the next most frequently occurring concept, approximately 12% of the items overall, and was also included in all five instruments. Four other ICF two-digit areas comprised at least 5% of all items, but none was included in all five instruments.

Figure 1B shows the same analysis for the COPD-specific instruments. Sensory functions/pain (~50% overall) and mental/emotional (~20% overall) were the two most common concept areas used and were included in all these COPD instruments.

Mobility and domestic life each comprised at least 5% overall but were not included in each of the five COPD instruments.

The most common core concept in the five head and neck cancer-specific instruments (Fig. 1C) was the voice and speech functions category, occurring in approximately 30% of the items. This category was included in all five instruments. Mental/emotional, sensory functions/pain, and digestive/metabolic/endocrine functions also achieved at least 5% overall but were not included in all these head and neck cancer instruments.

Mental/emotional and community/social/civic life were the most common core concepts, each more than 20% overall, in the five skin disease–specific instruments (Fig. 1D). Interpersonal interactions and self-care were also included in all five instruments. Three other areas also comprised at least 5% of the items, overall.

Table 6 shows predominant response options in the 20 disease-specific instruments in the four diagnostic categories with multiple instruments. The majority of the instruments (70%) used the same number of response options across all their items. Five-point scales were most commonly used (eight instruments), but overall six different numbers of response options were used, ranging from 2-point (yes/no, true/false) to 11-point (0–10 range) scales. Every disease area, however, showed variation across instruments in the type of response scale used. The Fibromyalgia Impact Questionnaire used a four-point Likert-type response and scaling approach for all but 2 of the 13 items.

Information on recall period is shown in Table 7. Most instruments (85%) used a single recall period for all their items. “The past week” was the most commonly used, particularly in arthritis and head and neck cancer. “Now” or “past month” or “past 4 weeks” were also very common. Every disease area, however, showed variation across instruments in the recall period used.

Table 2A,D presents the average-high-low percentages for the evaluation-aspect grammar components of items in the five instruments in the four disease areas. For the arthritis instruments, questions about difficulty with activities were the most common across instruments (~53% overall), with impact, severity, and frequency each comprising 7% to 15% overall. The most common evaluation aspects used in COPD instruments were extent (~39% overall) frequency (~26% overall), and severity, difficulty, and yes/no responses comprising 8% to 17% of all responses. Similarly, extent (~37% overall) and frequency (~28% overall) were the most commonly used in head and neck cancer instruments; bother and difficulty were also used (9% to 17%). For the psoriasis/skin disease instruments, extent (~38% overall) was most common; stress, frequency, and yes/no responses were the only other types used; in this disease area, four out of five instruments used only one aspect type.

Other aspect types not noted above but found across instruments and disease areas included satisfaction, problem, embarrassed, wellness, limitation, troubled, agreement, depressed, global rating of health, perception, and worry. In this grammar component as well, instruments take quite different approaches to evaluation aspects within disease areas.

Discussion

This article proposes a descriptive, rather than a prescriptive, grammar for evaluating conceptual content in COA instruments. At the item level, concepts consist of three components: a core concept consisting of an event, which may have a modifying clause (a modifier); an evaluation (both aspect and rating); and a recall period. The evaluation and recall period components largely represent the nature of the response to the core concept being evaluated. Together, the three components identify an item-level concept for use in evaluating treatment benefit.
<table>
<thead>
<tr>
<th>Main ICF component</th>
<th>Core concept (chapter)</th>
<th>ICF chapter code</th>
<th>Disease area (number of instruments)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Arthritis (5)</td>
<td></td>
</tr>
<tr>
<td>Body Functions (b)</td>
<td>Mental functions</td>
<td>b1</td>
<td>3 4 3 1 5 2 18</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sensory functions and pain</td>
<td>b2</td>
<td>4 4 0 1 1 2 12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Voice/speech</td>
<td>b3</td>
<td>0 5 0 0 0 0 5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Respiratory function</td>
<td>b4</td>
<td>0 3 1 1 0 0 5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Digestive, metabolic, and endocrine functions</td>
<td>b5</td>
<td>0 5 0 0 0 0 5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Intimacy</td>
<td>b6</td>
<td>1 1 1 0 0 0 3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Neuromusculoskeletal function</td>
<td>b7</td>
<td>2 1 0 1 0 0 4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Skin and related structure</td>
<td>b8</td>
<td>0 0 0 3 0 0 3</td>
<td></td>
</tr>
<tr>
<td>Body Structures (s)</td>
<td>Head and neck shape</td>
<td>s7</td>
<td>0 1 0 0 0 0 1</td>
<td></td>
</tr>
<tr>
<td>Activities and Particiation (d)</td>
<td>Tasks</td>
<td>d2</td>
<td>1 0 1 0 0 0 2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Communication</td>
<td>d3</td>
<td>0 1 0 0 0 0 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mobility</td>
<td>d4</td>
<td>5 0 3 1 0 2 11</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Self-care</td>
<td>d5</td>
<td>5 1 2 0 4 1 12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Domestic life</td>
<td>d6</td>
<td>4 0 2 1 0 2 9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Interpersonal interactions</td>
<td>d7</td>
<td>2 1 0 0 4 0 3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Major life areas</td>
<td>d8</td>
<td>2 3 0 1 0 0 6</td>
<td></td>
</tr>
<tr>
<td>Environmental Factors (e)</td>
<td>Personal consumption</td>
<td>e1</td>
<td>1 2 0 0 0 0 3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Family</td>
<td>e3</td>
<td>0 2 0 0 0 0 2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Social support</td>
<td>e4</td>
<td>1 1 0 0 0 0 2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Global</td>
<td>No ICF code</td>
<td>1 2 1 0 0 1 4</td>
<td></td>
</tr>
</tbody>
</table>

ICF, International Classification of Functioning, Disability, and Health; NA, not applicable/available.
Development of a descriptive grammar, rather than specific rules, for understanding the conceptual content of an item seems to be an appropriate approach for identifying item content. As in everyday speech, the grammar may be important for understanding an item’s content and patient’s response within the context of treatment benefit; this, in turn, may affect content validity and inform comparisons of results across instruments.

![Graph A](image1)

![Graph B](image2)

![Graph C](image3)

![Graph D](image4)

Fig. 1 – The average percentage of items within an instrument in given two-digit International Classification of Functioning, Disability, and Health (ICF) categories, as well as the high and low percentages, across the five instruments in the disease category. (A) Percentage of items by ICF category for five arthritis instruments. Other includes ICF chapters b3, b4, b5, b6, d3, d8, d9, e1, e3, e4, and globals (see Table 5). (B) Percentage of items by ICF category for five COPD instruments. Other includes ICF chapters b6, d2, d9, and globals (see Table 5). (C) Percentage of items by ICF category for five head and neck cancer instruments. Other includes ICF chapters b4, b6, b7, d3, d5, d7, d8, d9, s7, e1, e3, and e4 (see Table 5). (D) Percentage of items by ICF category for five psoriasis/skin disease instruments. Ave, average; COPD, chronic obstructive pulmonary disease; endo, endocrine; fctns, functions; metab, metabolic; pct, percentage.

Table 6 – Predominant number of response options, by disease and instrument.

<table>
<thead>
<tr>
<th>Number of response options</th>
<th>Arthritis</th>
<th>COPD</th>
<th>Head &amp; neck cancer</th>
<th>Psoriasis</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 patients</td>
<td>HAQ</td>
<td>SGRQ*</td>
<td></td>
<td>PSORIQoL</td>
</tr>
<tr>
<td>4 patients</td>
<td>AIMS2*, DASH, KOOS</td>
<td>EXACT*, PFSS*, LCADL, CRQ</td>
<td>DLQI-10*, PLSI, Skindex-29</td>
<td></td>
</tr>
<tr>
<td>5 patients</td>
<td></td>
<td>FACT H&amp;N, HNQOL, VHI</td>
<td></td>
<td>Skindex-29</td>
</tr>
<tr>
<td>6 patients</td>
<td></td>
<td>PDI-15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 patients</td>
<td>OAKHQOL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11 patients</td>
<td></td>
<td>QOT-RTI*, XQ</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

AIMS2, Arthritis Impact Measurement Scales 2; COPD, chronic obstructive pulmonary disease; CRQ, Chronic Respiratory Questionnaire; DASH, Disabilities of the Arm, Shoulder, and Hand; DLQI-10, Dermatology Life Quality Index 10; EXACT, EXacerbations of Chronic Obstructive Pulmonary Disease; FACT-H&N, Functional Assessment of Cancer for Patients with Head and Neck Cancer; HAQ, Health Assessment Questionnaire; HNQOL, Head and Neck Quality of Life Instrument; KOOS, Knee Osteoarthritis Outcome Scale; LCADL, London Chest Activities of Daily Living Scale; OAKHQOL, Osteoarthritis Knee and Hip Quality of Life Scale; PDI-15, Psoriasis Disability Index 15; PFSS, Pulmonary Functional Status Scale; PLSI, Psoriasis Life Stress Inventory; PSORIQoL, Psoriasis Index of Quality of Life; QOL-RTI, Quality of Life-Radiation Therapy Instrument; SGRQ, St. George's Respiratory Questionnaire; Skindex-29, Skindex-29; VHI, Voice Handicap Index; XQ, Xerostoma Questionnaire. * Used a different number of response options in a minority of items.
For comparison, analysis, and communication, however, core concepts are best understood if there is a cohesive taxonomy that is associated with the proposed grammar. The ICF is one such candidate but, while quite detailed for physical functioning, is relatively weak on emotional content. In addition, the ICF has yet to incorporate a diversity of formats for response options and recall periods into its structure. These limitations preclude its adoption for a complete PRO Concept Taxonomy that needs to include the full range of core concepts used in COAs as well as the response options and recall periods.

Table 7 – Predominant recall period used by disease and instrument

<table>
<thead>
<tr>
<th>Recall period</th>
<th>Arthritis</th>
<th>COPD</th>
<th>Head &amp; neck cancer</th>
<th>Psoriasis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Now/implied present</td>
<td></td>
<td>EXACT, PFSS*, SGRQ</td>
<td></td>
<td>PSORIqol</td>
</tr>
<tr>
<td>Last few days</td>
<td></td>
<td>LCADL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Past week</td>
<td>DASH, HAQ, KOOS</td>
<td></td>
<td>FACT H&amp;N, QOL-KTI, VHI, XQ</td>
<td>DLQI-10</td>
</tr>
<tr>
<td>Past 2 wk</td>
<td></td>
<td>CRQ</td>
<td></td>
<td>DLQI-10</td>
</tr>
<tr>
<td>Past 4 wk or month</td>
<td>AIMS2*, OAKHQOL</td>
<td></td>
<td>HNQOL*</td>
<td>PDI-15, PLSI, Skindex-29</td>
</tr>
</tbody>
</table>

AIMS2, Arthritis Impact Measurement Scales 2; COPD, chronic obstructive pulmonary disease; CRQ, Chronic Respiratory Questionnaire; DASH, Disabilities of the Arm, Shoulder, and Hand; DLQI-10, Dermatology Life Quality Index 10; EXACT, EXAcerbations of Chronic Obstructive Pulmonary Disease; FACT-H&N, Functional Assessment of Cancer for Patients with Head and Neck Cancer; HAQ, Health Assessment Questionnaire; HNQOL, Head and Neck Quality of Life Instrument; KOOS, Knee Osteoarthritis Outcome Scale; LCADL, London Chest Activities of Daily Living Scale; OAKHQOL, Osteoarthritis Knee and Hip Quality of Life; PDI-15, Psoriasis Disability Index 15; PFSS, Pulmonary Functional Status Scale; PLSI, Psoriasis Life Stress Inventory; PSORIqol, Psoriasis Index of Quality of Life; QOL-KTI, Quality of Life-Radiation Therapy Instrument; SGRQ, St. George’s Respiratory Questionnaire; Skindex-29, Skindex-29; VHI, Voice Handicap Index; XQ, Xerostoma Questionnaire.

* Used a different recall period in a minority of items.

Fig. 2 – Average percentage of items within an instrument with given evaluation aspects (i.e., the nature of the core concept being evaluated), as well as the high and low percentages, across the five instruments in the disease category. (A) Percentage of items by evaluation aspect for five arthritis instruments. *Other includes aspects such as satisfaction, wellness, extent, limitation, worry, and many others. (B) Percentage of items by evaluation aspect for five COPD instruments. *Other includes aspects such as satisfaction, distance, problem, and duration. (C) Percentage of items by evaluation aspect for five head and neck cancer instruments. *Other includes aspects such as satisfaction, embarrassed, perception, severity, and yes/no. (D) Percentage of items by evaluation aspect for five psoriasis/skin disease instruments. Ave, average; COPD, chronic obstructive pulmonary disease; pct, percentage.
In examining item content for the 21 disease-specific instruments analyzed, we found expected differences across diseases in terms of events, evaluation components, and duration as well as similarities within diseases. The analysis also indicated considerable variation within diseases in all structural parts. Thus, even if the events are similar across instruments within a disease, the variation in how those concepts are combined with evaluation and duration components could result in different concepts related to treatment benefit. For example, a COA measure with a core concept of pain and a recall period framed in terms of hours will generally be more meaningful for an acute migraine treatment than for one for a chronic condition such as arthritis. While the findings presented here focus on disease-specific instruments, the grammar was also found to be applicable to generic instruments, adding support for validity and generalizability of the proposed grammar. The validity information will be enhanced with the evaluation of more instruments. The ability of the grammar to fit with the ICF categories is another indication of its validity. Another planned approach to validation is to compare the item content identified using the grammar with that using the linking rules developed for use with the ICF [55–57].

In addition to introducing the grammar and taxonomy for evaluating the conceptual content in COA instruments, this analysis suggests their value for use in evaluating instrument selection within the context of a given application. For example, these tools can provide standards for validating concepts included in instruments being considered for use, for example, in the drug-approval process; these standards apply to both newly developed and extant instruments. Similarities of concepts at all levels can be interpreted as confirmatory information. Observed differences, for example, inclusion of a new singular or compound concept, would likely require exploratory analyses to determine whether these differences were due to problems in the instrument development process, such as mis-specification of the intended patient population, or due to changes in the nature of the disease impact due to technological advances.

Toward the goal of improving communication between researchers, sponsors, and regulators, all of whom bring their own professional languages to the discussion, we have proposed a simple grammar for evaluating and understanding item content. The grammar is designed to be used across COA instruments regardless of their theoretical foundations. This systematic approach to evaluating and classifying conceptually similar concepts in COA instruments aims to link patients’ assessments of treatment benefit directly to health decision making whether at the clinical practice level or the macro-level policy level.

Source of financial support: The authors have no other financial relationships to disclose.

REFERENCES


