
Pennifer Erickson, PhD,1 Richard Willke, PhD,2 Laurie Burke, RPh, MPH3

1OLGA, State College, PA, USA; 2Global Outcomes Research, Pfizer Inc., Peapack, NJ, USA; 3Study Endpoints and Labeling, Office of New Drugs, Center for Drug Evaluation and Research, Food and Drug Administration, Silver Spring, MD, USA

ABSTRACT

Objective: To facilitate development and evaluation of a PRO instrument conceptual framework, we propose two tools—a PRO concept taxonomy and a PRO instrument hierarchy. FDA’s draft guidance on patient reported outcome (PRO) measures states that a clear description of the conceptual framework of an instrument is useful for evaluating its adequacy to support a treatment benefit claim for use in product labeling. This guidance, however, does not propose tools for establishing or evaluating a PRO instrument’s conceptual framework.

Methods: We draw from our review of PRO concepts and instruments that appear in prescription drug labeling approved in the United States from 1997 to 2007.

Results: We propose taxonomy terms that define relationships between PRO concepts, including “family,” “compound concept,” and “singular concept.” Based on the range of complexity represented by the concepts, as defined by the taxonomy, we propose nine instrument orders for PRO measurement. The nine orders range from individual event counts to multiitem, multiscale instruments.

Conclusion: This analysis of PRO concepts and instruments illustrates that the taxonomy and hierarchy are applicable to PRO concepts across a wide range of therapeutic areas and provide a basis for defining the instrument conceptual framework complexity. Although the utility of these tools in the drug development, review, and approval processes has not yet been demonstrated, these tools could be useful to improve communication and enhance efficiency in the instrument development and review process.

Keywords: classification system, conceptual framework, patient-reported outcomes, PRO concept taxonomy, PRO instrument hierarchy.

Introduction

The 2006 Food and Drug Administration draft guidance on patient-reported outcome (PRO) measures states that one of the first steps in the instrument selection or development process is the identification of the conceptual framework of each instrument [1]. The framework specifies the purpose for each item in terms of the instrument’s measurement goal and specifies how each item is to be used, either as a single-item concept or grouped together to form more complex concepts scored according to the instrument’s measurement structure and scoring system. The instrument can be deemed adequate to support a targeted statement of treatment benefit (i.e., claim) if the instrument measures the claimed concept in a well-defined and reliable way. By recommending the specification of the conceptual framework for each instrument, FDA recognizes the extensive variation that exists among PRO instruments. The tools described here offer a systematic approach to establishing and evaluating any instrument’s conceptual framework.

Instruments used in clinical research studies are known to differ in content depending on their intended application, for example, diagnosis, disease severity, and patient characteristics. These factors, in turn, determine the most relevant concepts for measuring treatment impact. Instruments may also differ according to developers’ perspectives on how to represent PRO concepts and their relationships; for example, researchers trained in medicine, psychology, and economics have developed instruments with different item formats, content, measurement structures, and scoring systems [2–5]. Reviews of compendia of health status and well-being measures present a more complete perspective of the diversity of concepts and measurement structures used in generating scoring systems for measures used in various fields, including pharmacoeconomics, health services research, geriatrics, mental health, and nursing [6–11].

Within the PRO field, researchers, including Fries, Guyatt, and Spilker [12–14], have proposed taxonomies for classifying health-related quality-of-life (HRQoL) concepts; these systems, however, have not as yet been operationalized. Existing classification operational systems, such as the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV), International Statistical Classification of Diseases and Related Health Problems (ICD), and International Classification of Functioning, Disability, and Health (ICF), [15–19], illustrate the clustering of concepts and diagnoses and their hierarchical arrangement into concepts of increasing complexity. These, however, have been designed for enumeration and epidemiologic analysis rather than for the type of evaluative decision-making required in the drug approval process. Our review of labeling approved by FDA indicated that PRO instruments of different complexities, from single items of event counts to multitem, multiconcept instruments have been used to support claims of treatment benefit [20]. Furthermore, this review suggested that it would be possible to link an instrument’s content and measurement structure to the nature of a statement of treatment benefit. That is, there is an interrelationship between the intended claim and the measure that supports it.

The ability to identify and codify this relationship has several advantages to sponsors, regulators, as well as to outcomes researchers more broadly. First, the sponsor and FDA need to understand the complexity of the concept in the desired claim because it will determine the adequacy of the instrument used to...
support that claim. From FDA’s point of view, more complex claims are likely to require more comprehensive instruments that have been demonstrated to capture all the important aspects of the complex concept in the targeted patient population [21]. Second, matching the complexity of the claim to patients’ and physicians’ perspectives of disease burden and impact can be important to the external credibility and effect of the claim. Third, being able to link a PRO instrument explicitly to regulatory or clinical decision-making via the conceptual framework can be both a rewarding and challenging aspect of study design and implementation. Moreover, specification of the relationship between a statement of treatment benefit and the PRO instrument that supports this claim incorporates the need for using standard, well-established psychometric methods to demonstrate properties, such as content and construct validity, as integral components of the decision-making process.

To set forth a systematic method for depicting an instrument’s conceptual framework, this article proposes a “PRO Concept Taxonomy” and a “PRO Instrument Hierarchy.” These two tools endeavor to resolve inconsistency and confusion when conceptualizing and quantifying treatment benefit measured by PRO instruments. The PRO Concept Taxonomy incorporates key terms, including “singular” concept, “compound” concept, and “family” concept; usage of these terms is proposed as a way of adding clarity to the development of an instrument’s conceptual framework. This proposed classification system is generalizable across a wide range of families and concepts.

The PRO Instrument Hierarchy connects the conceptual content of a PRO instrument that has been selected to support the intended claim with the instrument’s measurement structure and scoring system, thereby completing the description of the instrument’s conceptual framework. By linking the claim made with the complexity of the instrument used to support it, we can plan a measurement strategy for future labeling goals.

Methods for Developing the PRO Concept Taxonomy and PRO Instrument Hierarchy

The first step in developing the taxonomy and hierarchy was to evaluate PRO concepts that were identified in our review of the Clinical Studies sections of the labeling for 215 new products approved in the United States from January 1997 through December 2002 [20]; labeling for 64 of these products was found to report at least one PRO. We attempted to identify the actual PRO instrument used to measure the PRO concept and each instrument was evaluated in terms of its conceptual framework to determine the instrument’s relationship to the PRO concept identified. In this article, we use the term “concept” to refer to an aspect of how patients feel or function that is expressed qualitatively; when measured by a PRO instrument, a concept is represented by items and domains.

The second step was to validate the taxonomy and hierarchy by evaluating the labeling for the 142 new products approved by FDA from January 2003 through December 2007; labeling for 36 products reported at least one PRO. The PRO concepts and instruments found in labeling for the 1997–2007 period can be found at: http://www.ispor.org/Publications/value/VI/Hsupplementary/VG1218_Erickson.asp. The same methods were used for this review as for that of the 1997–2002 labeling. Third, we broadened the scope of our evaluation of PRO instruments to include formal scales beyond those that appeared in the new product labeling using information from the On-Line Guide to Quality-of-Life Assessment (OLGA) [6,22]. OLGA’s comprehensive database includes information on thousands of instruments that are of potential relevance for supporting a claim of treatment benefit. Based on selection criteria designed to identify instruments of diverse conceptual content and measurement structures, the conceptual frameworks of 25 instruments were formally evaluated. This step provided assurance that the taxonomy and hierarchy would be relevant not only to instruments used in previous labeling, but also to those that might appear after 2007.

These evaluations indicated that to fully understand the concept, or concepts, measured by a single instrument or battery of instruments, it is necessary to understand the relationships between the included concepts within the context of their use in the intended claim. For example, a claim of treatment benefit for a new migraine product is commonly stated in terms of five separate symptoms (defined below). Because there is no explicit specification of an interrelationship between them, five symptom-specific instruments are used to provide an implicit, rather than a measured, statement about treatment impact of the more general concept of migraine symptoms.

On the other hand, arthritis-related physical function is frequently expressed in terms of abilities to perform everyday activities, such as basic activities of daily living (ADLs) and instrumental ADLs (IADLs), for example, shopping, managing money, doing heavy housework, and mobility. When the relationships between the general and specific concepts is explicitly recognized, they can be measured using a single instrument, such as the Health Assessment Questionnaire Disability Index (HAQ-DI) [23], and the obtained scores can provide explicit information about treatment impact on both the more general concept as well as the specific abilities.

The PRO Concept Taxonomy

As a result of our evaluation of instruments, we define four nested levels of concepts that represent a practical limitation on the number of levels relevant for making meaningful statements about treatment benefit using PROs, a fifth level we define as concepts that are too basic for supporting meaningful claims (see Fig. 1). Concepts in lower levels of the nested arrangement are more specific than those in the higher levels. Understanding relationships between concepts enables researchers to apply an instrument that is appropriate for the purpose of measurement.

To facilitate a systematic method for depicting a conceptual framework, we define three terms: “family,” “compound concept,” and “singular concept.” A family is a taxonomic category that consists of subcategories, much like species and subspecies in biology. In the PRO context, families can be thought of as higher-level concepts that have subconcepts consisting of compound and singular concepts.

Families may be either generic or specific with respect to disease or condition. Generic families, such as mental, physical, and social function [24–26] are too general for meaningful, product-related discussions and measurement. Specific families, on the other hand, categorize concepts that are related to key diagnostic and therapeutic aspects and, thus, are useful for discussing treatment benefit; each specific family can be placed within a generic family. For example, the specific family of migraine symptoms, which is traditionally defined in terms of nausea, vomiting, pain, phonophobia, and photophobia, is located within the generic family of signs and symptoms.

Each family, whether generic or specific, comprises at least one singular concept that both patients and their health-care decision-makers could consider to be a meaningful goal of treatment benefit, for example, pain intensity. Singular concepts may have low-level singular concepts that are considered to be too basic for use in labeling, for example, ability to cut meat. A
compound concept is defined as consisting of at least two singular concepts; for example, the concept “basic activities of daily living” typically includes bathing, toileting, transferring, and dressing.

The PRO Concept Taxonomy is intended to provide structure to the task of establishing and reviewing a conceptual framework. This task requires identification of the concepts represented by instrument scores, identifying all items that contribute to that score, and diagramming the nesting of concepts within one or more families where appropriate, as illustrated in Figure 2. Singular concepts, and low-level singular concepts, are the most fundamental units in the taxonomy and can be considered as the “building blocks” of compound concepts. A compound concept may be made up of two types of singular concepts: 1) those that include low-level singular concepts, as shown in Family 1; and 2) those that can be measured with one item, as shown in Family 2. The type and number of these singular concepts depends on the disease and its treatment as well as the compound concept that represents the goal of measurement and corresponds to the labeling targets. A statement of benefit may be based on information about a single family or multiple families.

Figure 1 Patient-reported outcome concept taxonomy: depicts relationships between concepts. ADL, activity of daily living.

Figure 2 Patient-reported outcome instrument hierarchy: depicts concepts, measurement structure, and relationship to hierarchy.
A PRO Concept Taxonomy and Instrument Hierarchy

As shown in this figure, an aggregate is a compound concept that explicitly includes multiple families, for example, HRQoL. A global concept includes one or more families that are implicitly defined and aggregated by the patient, for example, self-rating of health, and is outside the scope of a classification system that is based on clearly identified concepts and their explicit relationships.

The main organizing unit for specifying one or more concepts is the family. Each concept must belong to one family and, conversely, each family can have few or many singular concepts. In fact, a very simple depiction of the PRO Concept Taxonomy can contain one singular concept within a single family in a given application, for example, arthritis-specific pain within the HAQ. More complex, single-family concepts may have low-level singular concepts that are used to form singular concepts. Singular concepts may be used to form compound concepts if the instrument development process provides empirical evidence that the compound concept is defined by the singular concepts.

Procedures for identifying PRO concepts and their relationships are referenced in the FDA draft PRO guidance and documented in other publications [27–31]. These established methods reflect the importance of using both qualitative and quantitative techniques to assure that an instrument provides a suitable measure of the intended measurement goal. Instruments developed using such procedures are most likely to contain items and domains that adequately represent the concepts that are meaningful to both patient and health-care professional, and to incorporate an approach to measurement that creates scores appropriate for the intended use, for example, as clinical trial end points.

Consideration of these PRO Concept Taxonomy principles can assist in depicting an instrument’s conceptual framework. By comparing an instrument’s taxonomic structure with a product’s targeted labeling claims, the adequacy or an instrument can be assessed and researchers can gain insights into the additional instrument development work needed to support those claims. Insight into the complexity of a concept can also be useful when designing studies to support claims related to that concept.

The PRO Instrument Hierarchy

The second step in specifying an instrument’s conceptual framework is to formalize relationships between concepts through the identification of the measurement structure and scoring system and verify this against the measurement goals and the targeted claim. Our review of approved labeling indicated that, regardless of taxonomic structure, instruments could be grouped into nine categories, representing increasing orders of conceptual and measurement complexity. Table 1 shows the nine orders in the hierarchy in terms of their number of families and concepts, and measurement structure, along with examples to illustrate the type of PRO instrument in each order. As indicated in columns 2 and 3, multiple-family instruments may be made up of singular or compound concepts within the individual families. The number and type of families and concepts within an instrument varies depending on the intended use of the instrument. Some instruments with multiple families may also permit the formation of an aggregate concept that may support a claim of “health-related quality of life” (HRQoL) if the included concepts meet the FDA’s HRQoL definition [1]. A multifamily instrument may have a validated measurement structure that permits it to support end points of more than one order, depending on the concepts chosen as study end points (e.g., the 36-Item Short-Form Health Survey [SF-36]; see below).

Order 0 categorizes the simplest type of conceptual framework and Order 8 categorizes the most complex. All PRO instruments, whether generic, disease specific, treatment specific, or global, belong to at least one family and thus can be placed in at least one order in this hierarchy. Each order is also characterized by a measurement structure that indicates the degree to which scores for singular concepts can be combined to form higher-level scores. Thus, each instrument score (or set of scores) becomes a study end point and the concept represented by that score, or set of scores, determines the particular order in the hierarchy that score is assigned. PRO measures that are based on patient reports of frequencies or occurrences of disease- or treatment-related events are classified in Order 0. Instruments that record patients’ evaluative responses, for example, severity or bothersomeness, about symptoms, functions, or perceptions are placed in Orders 1–8.

Measures that assess a frequency count as a singular concept in one family, such as the number of stools observed in the past week, are classified into Order 0 and support very specific statements about treatment effect. Instruments that elicit a patient’s evaluation of a singular concept in one family are classified into Order 1; like instruments in Order 0, these also support very specific statements about treatment effect. The measure of oculus itching in ALAMAST labeling is an example of an Order 1 instrument.

Global item measures are placed in Order 2 as each assesses a compound, rather than a singular, concept. Global item measures provide general information that is difficult to use as the only evidence to support a clinical decision. They are included in the PRO Instrument Hierarchy, however, as they have frequently appeared in labeling, especially those for treatments of rheumatoid arthritis.

PRO measures in Order 3 assess singular concepts within one family measured as a battery. Order 3 instruments differ from those in Order 1 in that the singular concepts are clustered together in labeling in some explicit way, such as in the measurement of “time to symptom improvement” or in the need to “win” simultaneously on a cluster of symptoms. The battery of instruments measuring four migraine symptoms in IMITREX labeling (Table 1) is an example of an Order 3 measure. These measures support symptom-specific statements of treatment benefit and when taken into consideration altogether implicitly demonstrate, rather than explicitly measure, treatment benefit at the family level (e.g., migraine symptoms).

Order 4 measures support statements of treatment benefit based on both the singular concepts and the family, as illustrated by the excerpt from ARAVA labeling in Table 1. Instruments in Order 4 have a measurement structure that provides a profile of scores that allows for meaningful interpretation when comparing scores across domains throughout the duration of treatment. Order 5 measures have four levels within one family and can support statements of treatment benefit at three levels, namely, the singular and compound concept as well as the family levels. Although no instruments of this type were found in our review of approved labeling (see below), we include it for completeness.

Orders 6–8 instruments include two or more families with two or more concepts. Like Order 4 instruments, these instruments also generate profiles of scores that can support measurement of concepts at various levels and offer multiple study design and analysis options. Order 6 instruments, like those in Order 3, measure individual concepts, but unlike Order 3, the concepts in Order 6 instruments have a measurement approach, for example, summated ratings, that allows for comparisons between the family concepts; the SF-36 profile is an example of an Order 6 instrument [32]. Order 7 measures combine multiple singular or compound concepts into families or an aggregate that includes
Table 1 PRO instrument hierarchy for classifying PRO instruments according to their taxonomic and measurement structures, with examples of PRO instruments and statements of treatment benefit from existing prescription drug labeling

<table>
<thead>
<tr>
<th>Order number</th>
<th>Families*</th>
<th>Concepts*</th>
<th>Taxonomic and measurement structure</th>
<th>Claim(s) supported by instrument</th>
<th>With example of PRO statement of treatment benefit*</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>I</td>
<td>1 S</td>
<td>I or more items in a singular concept that assess frequencies or occurrences that are disease or treatment related. Example: Number of stools per week</td>
<td>Specific claim about the reported event</td>
<td>“Patients on ZELNORM also experienced an increase in median number of stools from 3.8/week to 6.3/week at month 1 ...”</td>
</tr>
<tr>
<td>1</td>
<td>I</td>
<td>1 S</td>
<td>I or more items eliciting patient evaluation of either a symptom, function, or perception. Example: Occular itching</td>
<td>Specific claim about the evaluated singular concept</td>
<td>“ALAMAST was significantly more effective than placebo after 28 days in preventing ocular itching associated with allergic conjunctivitis.”</td>
</tr>
<tr>
<td>2</td>
<td>I+</td>
<td>1 C</td>
<td>A global, compound concept measured by a single item. Example: Overall rating of the condition of dry mouth now compared with before starting treatment.</td>
<td>General claim that reflects the content of the item</td>
<td>“Statistically significant global improvement in the symptoms of dry mouth was seen ...” (EVDXAC)</td>
</tr>
<tr>
<td>3</td>
<td>I</td>
<td>2+ S</td>
<td>Multiple singular concepts representing a cluster of disease-related concepts with one or more measurement approaches that allow for individual concept scores. There is no family score. Example: Headache response defined in terms of severity of headache pain. Associated symptoms of nausea, photophobia, and phonophobia were also assessed.</td>
<td>Concept-specific claims but no family-level claim. There are as many claimable end points as there are concepts. Example: “The percentage of patients achieving headache response 2 and 4 hours after treatment was significantly greater among patients receiving IMITREX for patients with migraine-associated nausea, photophobia and/or phonophobia at baseline, there was a lower incidence of these symptoms at 2 hours (Study 1) and at 4 hours (Studies 2,4, and 3).”</td>
<td>Both concept-specific and family-level claims. There are at least three claimable end points. Example: “The mean change from baseline in functional ability as measured by the HAQ Disability Index (HAQ DI) in the 6 and 12 month placebo and active controlled trials is shown in Figure 4. ARAVA was statistically superior to placebo in improving physical function. Superiority to placebo was demonstrated consistently across all eight HAQ DI subcales (dressing, arising, eating, walking, hygiene, reach, grip and activities) in both placebo controlled studies.”</td>
</tr>
<tr>
<td>4</td>
<td>I</td>
<td>2+ C</td>
<td>Singular concepts are expressed in 2+ singular concepts with a measurement approach that allows for a compound family score. Concept and family scores are measured using a scoring system that allows direct comparison of concepts. Example: Health Assessment Questionnaire Disability Index (HAQ DI)</td>
<td>Both concept-specific and family-level claims. There are at least three claimable end points. Example: “The mean change from baseline in functional ability as measured by the HAQ Disability Index (HAQ DI) in the 6 and 12 month placebo and active controlled trials is shown in Figure 4. ARAVA was statistically superior to placebo in improving physical function. Superiority to placebo was demonstrated consistently across all eight HAQ DI subcales (dressing, arising, eating, walking, hygiene, reach, grip and activities) in both placebo controlled studies.”</td>
<td>Both concept-specific and family-level claims. There are at least three claimable end points. Example: “The mean change from baseline in functional ability as measured by the HAQ Disability Index (HAQ DI) in the 6 and 12 month placebo and active controlled trials is shown in Figure 4. ARAVA was statistically superior to placebo in improving physical function. Superiority to placebo was demonstrated consistently across all eight HAQ DI subcales (dressing, arising, eating, walking, hygiene, reach, grip and activities) in both placebo controlled studies.”</td>
</tr>
<tr>
<td>5</td>
<td>I</td>
<td>1+ C and 1+ S</td>
<td>Compound concepts each have at least one subconcept with a measurement approach that allows for the calculation of subconcept and concept scores as well as a family score. Both concept and family scores represent compound concepts. Example: None found</td>
<td>One family, and concept and subconcept claims; there are as many claimable end points as there are end points in the three levels. Example: “None found in labeling 1997–2007”</td>
<td>“None found in labeling 1997–2007”</td>
</tr>
<tr>
<td>6</td>
<td>2+</td>
<td>2+ S</td>
<td>Multiple singular concepts, each of which represents a family, with a measurement approach that allows comparison across concepts. There is no aggregate score. Example: Walking Impairment Questionnaire (WHQ)</td>
<td>Concept-level claims. There are as many claimable end points as there are concepts. Example: “The Fullkirk Impairment Questionnaire assesses the impact of a therapeutic intervention on walking ability. In a pooled analysis, patients reported improvement in their walking speed and walking distance. (PLETAL).”</td>
<td>“The mean change from baseline in functional ability as measured by the HAQ Disability Index (HAQ DI) in the 6 and 12 month placebo and active controlled trials is shown in Figure 4. ARAVA was statistically superior to placebo in improving physical function. Superiority to placebo was demonstrated consistently across all eight HAQ DI subcales (dressing, arising, eating, walking, hygiene, reach, grip and activities) in both placebo controlled studies.”</td>
</tr>
<tr>
<td>7</td>
<td>2+</td>
<td>2+ S or C</td>
<td>Concepts are measured in terms of 2+ concepts and 2+ families with a measurement approach that allows calculation of concept and family scores that can be compared. There is an aggregate score that combines more than one family but omits at least one major family needed to support the HRQoL concept. Example: Asthma Quality of Life Questionnaire (AQLQ—Juniper)</td>
<td>Family and concept-level claims, with as many claims as there are families and concepts. Example: “The subjective impact of asthma on patient’s perception of health was evaluated through use of the AQLQ. Patients receiving ADAIR DISKUS 100/50 had clinically meaningful improvements in overall asthma-specific quality of life as defined by a difference between groups of at least 0.5 points in change from baseline.”</td>
<td>“The mean change from baseline in functional ability as measured by the HAQ Disability Index (HAQ DI) in the 6 and 12 month placebo and active controlled trials is shown in Figure 4. ARAVA was statistically superior to placebo in improving physical function. Superiority to placebo was demonstrated consistently across all eight HAQ DI subcales (dressing, arising, eating, walking, hygiene, reach, grip and activities) in both placebo controlled studies.”</td>
</tr>
<tr>
<td>8†</td>
<td>3+</td>
<td>3+ C</td>
<td>Family and concept scores measurement approach that allows comparison across families and concepts. There is an aggregate score that includes all families needed to support the HRQoL concept. Example: Sickness Impact Profile (SIP)</td>
<td>An overall (potentially HRQoL), as well as multiple family and concept claims; there are as many claimable end points as in the three levels plus the aggregate score. Example: “The SIP, a multiscale in 12 concepts designed to assess the patient’s functioning in multiple areas. Data for the overall SIP score at baseline and change from baseline at 3 months are presented in Table 2. For TASMAR, the change from baseline was statistically significant for the 200 mg tid treatment arm, with a p-value of 0.01.”</td>
<td>“The mean change from baseline in functional ability as measured by the HAQ Disability Index (HAQ DI) in the 6 and 12 month placebo and active controlled trials is shown in Figure 4. ARAVA was statistically superior to placebo in improving physical function. Superiority to placebo was demonstrated consistently across all eight HAQ DI subcales (dressing, arising, eating, walking, hygiene, reach, grip and activities) in both placebo controlled studies.”</td>
</tr>
</tbody>
</table>

C: compound concept; HRQoL: health-related quality of life; PRO: patient-reported outcome; S: singular concept. *Labeling statements are taken from: http://www.accessdata.fda.gov/scripts/cder/drugsatfda [46], the 2002 or the 2006 PDR [47]. **Any instrument or battery of instruments that provides an overall score without documentation that supports an underlying theoretical model or justification for combining multiple families of concepts should not present the overall score for decision-making. If such a score is used, a caveat about the lack of an appropriate measurement structure should be stated in a footnote. Note The examples in this table are drawn from the review of new prescription drug labeling approved between 1997 and 2002. These examples illustrate relationships between statements of treatment benefit and the measurement structures of various instruments. They do not, however, provide assurance that the same relationships will be applied to future drug approvals. “
more than one family; these instruments can be used to measure both the concepts and the families. In addition, Order 7 instruments may be used to measure the concept represented by the aggregate score.

Order 8 measures are the most “complex,” both conceptually and practically, because they: 1) measure three or more families, including all families needed to support the HRQoL concept as specified in FDA’s draft guidance, i.e., physical, psychological/emotional, and social functioning; 2) have multiple domain scores; and 3) incorporate measurement approaches that support the calculation of an aggregate score. Order 8 instruments can be used to measure singular concepts, family concepts, or aggregate concepts. A conclusion that a treatment impacts HRQoL would be based on an Order 8 instrument.

**Depicting the Conceptual Framework**

The conceptual framework of a battery of instruments proposed for evaluating the benefit of a new migraine treatment, that is, an Order 3 battery of instruments, is illustrated in Figure 3 using the taxonomy and hierarchy. The first step in developing this framework is to identify a set of signs and symptoms related to migraine headache that are recognized by patients and clinicians as being meaningful for defining migraine treatment response. The resulting specific family of migraine symptoms is represented by a cluster of five singular concepts, shown as the taxonomic structure in Figure 3. The dashed lines connecting the singular concepts to the family level indicate that relationships between the individual symptoms and the family, the measurement structure, are implied rather than explicit, that is, the scoring system for the five symptoms does not include a combined symptom score at the family level. In this example, a conclusion concerning a treatment benefit (migraine response) would be based on improvement in every symptom depicted in the conceptual framework.

Figure 4 shows the use of the taxonomy and hierarchy to depict the conceptual framework of the HAQ-DI for supporting labeling claims at both the family and compound concept levels, an instrument in Order 4. As shown in this figure, the HAQ-DI measures a specific family, defined by the eight singular concepts, which are, in turn, composed of low-level singular concepts. The solid lines indicate that the instrument’s measurement structure provides a rationale for combining low-level singular concepts to form explicit statements about patient performance of eight singular concepts as well as the compound concept of physical disability, which is expressed in a single score within the family of arthritis-related physical function.

In developing both the taxonomy and hierarchy, we started with the evaluation of a given instrument according to its content, measurement structure, and scoring system. This process produces a depiction of the conceptual framework as illustrated in Figures 3 and 4. The orders in the PRO Instrument Hierarchy also indicate the type of claim that the instrument can support.

**Evaluating the PRO Instrument Hierarchy Using Recently Approved Labeling**

The explicit relationships between the PRO instrument’s conceptual content, expressed in terms of the PRO Concept Taxonomy, and the treatment benefit statements in labeling, reflected in the PRO Instrument Hierarchy, were evaluated and validated in two separate stages. A previous analysis showed that labeling for 64 (30%) of the 215 new drugs approved from 1997 to 2002 included a treatment benefit statement (in the Clinical Studies section) about a concept measured by a PRO instrument [20]. We first reanalyzed the labeling for these 64 drugs, and classified the conceptual frameworks represented by the PRO statements therein into one of the nine categories described in Table 1. During this first stage, the PRO Instrument Hierarchy was adapted to better fit the actual labeling statements observed. To validate this hierarchy, we then analyzed the labeling of the 142 new drugs approved in 2003–2007 (following the same criteria used in the 1997–2002 study), of which 36 contained PRO-based
statements in their Clinical Studies section, to determine whether those statements and their implied conceptual frameworks mapped well into the hierarchy. This second mapping determined that no changes to the basic structure of the hierarchy were needed, but we felt it was appropriate to modify the description of Order 4, from “There are at least 3 claimable end points” to “This may allow 3 or more claimable end points.”

The percentage of times that each order occurred, for each of the two periods examined, is shown in Figure 5. Percentages add to more than 100% because the labeling for many drugs (38 of 100) contains more than one order of PRO statement. For some orders, the rate of use was similar between periods, in others it was not; some of the variation observed is due to differences in types of drugs approved between periods, as described below.

Simple event counts (Order 0) and singular PRO concepts measured with one or more items (Order 1) were the most commonly occurring orders, present in labeling for 40 and 52 of the 100 drugs, respectively. Some frequently used event counts

---

**Figure 4** Intended statement of treatment benefit using the HAQ-DI: improve physical disability in rheumatoid arthritis patients.

---

**Figure 5** Patient-reported outcome (PRO) instrument orders in new drug labeling*.
were cough (immunologic agents), partial seizure frequency (anti-epileptic agents), and use of rescue medications (antimigraine and respiratory agents). Frequently used singular PRO concepts were: pain intensity, symptom assessments (several areas), ocular itching (ophthalmics), and dyspnea (cardiovascular).

PRO concepts of Order 2 (global concepts), Order 3 (a cluster of singular concepts), and Order 4 (1 family represented by one compound concept containing 2+ singular concepts) were the next most common, appearing in labeling for 25, 18, and 16 different drugs, respectively. General concepts were most common for anti-inflammatory agents, as a patient global score is part of the American College of Rheumatology 20/50/70 criteria used in rheumatoid arthritis; these accounted for the labeling of 10 drugs out of the 26 with global scores [33,34]. Other statements classified as globals were: time spent in on-off states for Parkinson’s disease (five cases); ability to perform normal activities; and satisfaction with treatment. Interestingly, global items were rarely the only PRO concept in labeling (4 out of 26 cases). Global concepts were less common in labeling approved between 2003 and 2007, primarily due to only one drug for rheumatoid arthritis being approved during that period.

Order 3 PRO measures (which measure a cluster of singular concepts) were most common among gastrointestinal agents and antimigraine products, where different symptom concepts (e.g., phonophobia, photophobia, nausea) are clustered together as a single disease-specific family of concepts (migraine symptoms).

Use of Order 3 instruments was much higher in the 1997–2002 than in 2003–2007 due to the approval of 6 migraine drugs in the earlier period, all with Order 3 PRO measures, as opposed to no migraine drugs in the later period.

In the earlier period, all but one of the approvals based on Order 4 measures referenced the HAQ Disability Index (or M-HAQ) for anti-inflammatory products [23,33]; the only other Order 4 instrument was the total nasal and non-nasal symptom score for a respiratory product. In the later period, however, there was more varied use of Order 4, including the Erectile Function domain of the International Index of Erectile Dysfunction, the Functional Living Index—Emesis, the Sheehan Disability Scale, and the Alzheimer’s Disease Cooperative Study-Activities of Daily Living Inventory (ADCS-ADL) [36–39].

More complex PRO measures were less common, with no examples of Order 5 instruments occurring in this set of labeling, and a total of 21 examples with Order 6–8 measures. Order 6 measures (2+ families represented by 2+ singular concepts with a profile of scores) included the Walking Impairment Questionnaire (cardiovascular agents), the SF-36 profile of scores (diagnostics), the Quality of Life in Narcolepsy [40], the Toronto Western Spasmodic Torticollis Rating Scale (TWSTRS) [41] (central nervous system agents), and the International Index of Erectile Function [36] (urologic agents). The SF-36 profile of scores was the only Order 6 PRO measure during 2003–2008, used for the lone antiarthritis drug approved in the later period.

Examples of Order 7 PRO measures (2+ families represented by 2+ singular or compound concepts, allowing for family or aggregate scores) also primarily occurred for anti-inflammatory products, based on the SF-36 Physical Component Score (PCS) and Mental Component Score (MCS) scores or the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) composite score [42]; one was based on the total score of the Fibromyalgia Impact Questionnaire [43]. The two examples of an Order 8 PRO measure (3+ families, 3+ concepts, with an aggregate score) were the sickness impact profile [44] total score, used for an anti-Parkinsonian product, and a claim for “improved health-related quality of life” for treatment for paroxysmal nocturnal hemoglobinuria, based on results of the European Organization for Research and Treatment of Cancer Quality-of-Life Questionnaire-30-items (EORTC-QLQ-C30) [45].

In reviewing these results across orders and time periods, there has been relatively less frequent use of Orders 2–8 since 2003. Of those drugs approved in 1997–2002 with PROs in their labeling, 59% (38/64) included at least one PRO of Order 2–8, while in 2003–2007, 42% (15/36) included at least one PRO of Order 2–8. Much of this difference is due to the nature of the drugs approved during these periods—in the earlier period, 15 arthritis or migraine drugs were approved, all having these higher-order PROs, while in the latter period, only one arthritis or migraine drug was approved. Not including those drugs in this comparison results in 47% (23/49) labels from 1997–2002 with Order 2–8 PRO’s, and 40% (14/35) in 2003–2007.

This analysis indicates that instruments that have commonly been used in the drug approval process fit within the nine orders in the PRO Instrument Hierarchy, based on both an evaluation and a validation labeling sample. This finding provides evidence for the relevance of both the taxonomy and hierarchy in characterizing PRO instruments to be used in clinical trials. Most of the PRO data led to statements of treatment benefit within one family rather than multiple families, with over half being used to make narrow statements of treatment efficacy, that is, based on singular concepts that did not explicitly include a statement of family-level benefit.

Discussion

Specific terminology and the PRO Concept Taxonomy and PRO Instrument Hierarchy are proposed as approaches for more systematically establishing and evaluating conceptual frameworks for PRO instruments used in trials to assess clinical benefit. Beyond providing structures for characterizing PRO measures, they supply outcomes researchers with tools for evaluating and explaining an instrument’s conceptual framework within the context of a specific claim. With improved clarity of this structure, the linkage between the underlying diagnostic or conceptual terminology and the outcome of the health-care intervention becomes stronger and more transparent.

The drug-approval process is unique in that it explicitly links the use of a PRO instrument to medical decision-making through a statement of treatment benefit. The PRO Concept Taxonomy and PRO Instrument Hierarchy are proposed as structures for clarifying this linkage and for locating the use of well-established and relevant psychometric methods within this process. For example, use of these methods to demonstrate an instrument’s content validity within the context of the intended claim is part of the depiction of an instrument’s concept taxonomy. Similarly, depiction of an instrument’s measurement structure is determined by use of well-established quantitative psychometric methods which, in turn, locate the instrument within the PRO Instrument Hierarchy, thereby indicating its suitability for the intended claim.

The review of 1997–2002 new drug labeling illustrated that the PRO Instrument Hierarchy, incorporating the principles of the PRO Concept Taxonomy, is relevant across a wide range of both therapeutic products and the measures chosen to demonstrate their clinical benefit; this finding was confirmed by a subsequent review of 2003–2007 new drug labeling. For example, the predominance of the use of simple PRO instruments—event counts and singular concept PRO instruments (Orders 0 and 1)—along with global items and disease-specific, single-family PRO instruments (Orders 2 and 3) fits with the specific state-
ments about treatment benefit. Aside from the global PRO instruments, which are rarely used in isolation, the connection between the PRO instrument and the disease or its treatment is probably most transparent in these cases and the underlying conceptual framework of the instrument need not be complex.

Use of instruments with multiple concepts was much less common, particularly outside the antiinflammatory area, suggesting that establishing a clear relationship between treatment of a specific disease and broader PRO concepts can be more challenging, both in theory and in practice. Nevertheless, there are sufficient examples of measures with multiple concepts and families to indicate the relevance of the taxonomy and hierarchy and to establish the potential value of measures based on complex concepts. Use of the hierarchy along with the concept taxonomy, beyond simply allowing for a better understanding of the full spectrum of PRO statements allowed in labeling over this 11-year period, should assist in making the determination when to consider and justify the use of more comprehensive measures.

Characterizing PRO instruments in a standardized way may improve not only the communication between industry and its regulators but also within the research community more broadly. For example, abstracts of clinical studies frequently use terms such as pain, physical function, and HRQoL to describe measures that may represent any of the orders in the hierarchy. Unless the abstract specifically names the instruments used, the reviewer must locate the article to fully understand both the concepts being measured and the conceptual framework of the instrument in order to interpret the findings. Even within an article, the exact concept(s) measured may be incompletely documented, leading to misinterpretation of findings. More careful attention to the naming of concepts with consideration for the PRO Concept Taxonomy and PRO Instrument Hierarchy will help to clarify the results of clinical studies using PRO instruments.

The work presented here is limited in several ways. First, our approach has been heavily influenced by use of PRO’s in new drug labeling and hence may not be as applicable to other areas using PRO’s. Second, it has been based on retrospective evaluation of instruments and labeling; prospective use may, and is in fact likely to, generate new considerations that could affect the proposed taxonomy and hierarchy. Third, while we have acknowledged the important role of measurement science, especially that of content validity, in the developing a conceptual framework, we have yet to explicitly incorporate this work into our specification of the two tools. And, perhaps most importantly, our approach has not yet been used, to the best of our knowledge, in any interactions between sponsors and regulators, nor has it been explicitly endorsed by any regulatory agency.

Finally, the terminology, taxonomy, and hierarchy described above are proposed as a way of improving clarity and consistency when studies intended to evaluate therapeutic impact are conceived, developed, evaluated, and communicated. It draws both from the existing theoretical literature and from what has been observed in approved labeling and in the regulatory setting. Nevertheless, refinements and extensions to improve the taxonomy and hierarchy to meet future needs are both encouraged and expected. The overriding goal is to better incorporate the most relevant and interpretable PRO measures into drug development, drug labeling, and ultimately, patient care.

The authors would like to acknowledge the helpful comments of three anonymous reviewers as well as those from a number of colleagues who reviewed earlier drafts. The views expressed in this article are those of the authors and do not reflect the positions of either Pfizer Inc. or the Food and Drug Administration.

Source of financial support: None.

References


12 Fries JF. The hierarchy of quality-of-life assessment, the Health Assessment Questionnaire (HAQ), and issues mandating development of a toxicity index. Control Clin Trials 1991;12(Suppl. 4):106S–17S.


31 Leidy NK, Vernon M. Perspectives on patient-reported outcomes: content validity and qualitative research in a changing clinical trial environment. Pharmaco economics 2008;26:363–70.


