Response to the FDA Draft Guidance for Industry document:
Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims (Docket 2006D-0044)

Jakob B Bjorner, MD, PhD, Barbara Gandek, MS, Jason Cole, PhD, Mark Kosinski, MA, Gene Wallenstein, PhD, Milena Anatchkova, PhD, John E. Ware Jr., PhD

QualityMetric Inc. and Health Assessment Lab

We would like to express our appreciation to the FDA for this well-written document, which presents a strong and cohesive argument for the use of PRO tools in clinical research as well as useful guidance on the sound application of PRO assessment in clinical trials.

QualityMetric and Health Assessment Lab strongly support the overall goal of the Guidance to ensure that drug claims pertaining to PRO are backed up by sound measurement and research design. Since the draft guidance is likely to elicit numerous responses from the various stakeholders in PRO research, we would like to initially point out the places where we particularly strongly support the principles in the current version of the guidance. We will then provide some specific suggestions for modifications to various parts of the guidance and finally discuss a number of issues in more general terms. We are particularly enthusiastic about the following parts of the guidance:

*III.A. Why Use Patient-Reported Outcome Instruments in Medical Product Development? (*Lines 92-137).
This section presents a very well written description of the potential benefits of PRO assessment. We think this section is ideal as it is now.

Figure 1: The PRO Instrument Development and Modification Process and lines 183-192.

The wheel and spokes diagram presents a very nice framework for understanding PRO instrument development.

A. Development of the Conceptual Framework and Identification of the Intended Application (lines 194-210, 249-279, 460-467).

We agree with and support the recommendation to specify a conceptual framework and identify the concepts and domains to be measured based on this conceptual framework. The notion that the conceptual framework should be related to the claim sought is important and we would suggest that this aspect be given even more weight in the guidance – in line with the presentations at the Mayo Clinic meeting. We have some suggestions for improvement of the sections on single items and multi-domain instruments (lines 212-237, please see below), but we strongly support the intention of the FDA to evaluate the conceptual model by evaluating the relationships between individual items, domains, and the general concepts (lines 249-256), and to evaluate the intended application and populations for the PRO instrument.

Section C, 4a. Defining a minimum important difference (lines 537-568).

We enthusiastically support the distinction between *minimal important difference (MID)* as a benchmark for interpreting mean differences and the *definition of a responder* as pertaining to a change in an individual. We do not agree that an MID is usually specific to the population under study (please see below), but we agree that the use of a variety of methods to determining MID is generally helpful.

Specific suggestions for modifications to the guidance:

Line 153-156

"Some PRO instruments (e.g., health-related quality of life instruments) attempt to measure both the effectiveness and the side effects of treatment. PRO instruments that are used in clinical trials to support effectiveness claims should measure the adverse consequences of treatment separately from the effectiveness of treatment."

Suggested revision:

"Some PRO instruments (e.g., health-related quality of life instruments) attempt to measure both the effectiveness and the side effects of treatment. Clinical trials to support effectiveness claims should be designed to assess the adverse consequences of treatment separately from the effectiveness of treatment."

Motivation:

Separation of side effects from treatment effects is an important topic that extends beyond a particular PRO instrument. Steps to separate these effects should be taken in various parts of the study design (e.g., frequency of assessment of side effects, efficient reporting mechanism). Further, the original sentence could be misunderstood to mean that the separation of effectiveness from side effects could only be achieved by a PRO instrument measuring these separately and not by combining an instrument aimed at evaluating effectiveness with an instrument aimed at evaluating side effects.

Line 178-181

"When considering an instrument that has been modified from the original, the FDA generally plans to evaluate the modified instrument just as it would a new one. Therefore, in such instances, we encourage sponsors to document the original development processes, all modifications made, and updated assessments of its measurement properties."

Suggested revision:

"When considering an instrument that has been modified from the original, the requirements for documentation will depend on the extent and nature of the modification and the level of documentation of the original instrument. Therefore, in such instances, we encourage sponsors to document the original development processes, all modifications made, and updated assessments of its measurement properties."

Motivation:

The original statement appears to contradict section D line 582-583: *The extent of additional validation recommended depends on the type of modification made.* The extensive definition of modification in section D also renders excessive the plan to evaluate modified instruments just as new ones.

Line 214-223

"If the concept of interest is general (e.g., physical function), a single-item PRO instrument is usually unable to provide a complete understanding of the treatment's effect because a single item cannot capture all the domains of the general concept.
For this reason, single-item questions about general concepts that imply multiple domains rarely provide sufficient evidence to support claims about that general concept. However, single-item questions about general concepts can be useful to help interpret multi-item measures of the same concept and to determine whether important items or domains of a general concept are missing (e.g., when results using single general questions do not correlate with results using a multi-item questionnaire, this may be evidence that the questionnaire is not capturing all the important domains of the concept contained in the claim)."

Suggested revision:

"If the concept of interest is general (e.g., physical function), a single-item PRO instrument is usually unable to provide a complete understanding of the treatment's effect because a single item cannot capture all the domains of the general concept.
For this reason, single-item questions about general concepts that imply multiple domains rarely provide sufficient evidence to support claims about that general concept. However, single-item questions about general concepts can be valid measures of the patient's own overall assessment of the concept. Thus, care should be taken in the labeling of the concept."

Motivation:

The original text could be confusing: if a single-item measure does not capture all domains of a concept, it is hard to understand why a low correlation with results from a multi-item questionnaire should be taken as evidence that the questionnaire is not capturing all domains. The part of the multi-item questionnaire that shows low correlation with the single item could be the very part that is relevant but not captured by the single item. On the other hand, global single items can be valid measures of the patient's global assessment of a health domain. One example of this is the strong body of evidence supporting the validity of single items on general health perception (see e.g. [1;2])

Line 275-278

"The FDA plans to compare the patient population used in the PRO instrument development process to the study populations enrolled in clinical trials to determine whether the instrument is appropriate to that population with respect to patient age, sex, ethnic identity, and cognitive ability."

Suggested revision:

"The FDA plans to compare the patient population used in the PRO instrument development process to the study populations enrolled in clinical trials to determine whether the instrument is appropriate to that population with respect to pertinent demographic variables, such as patient age, sex, ethnic/racial identity, and cognitive ability. Pertinent demographic variables are those which have been empirically demonstrated to impact the construct measurement by the PRO. "

Motivation:

The largest change is to the inclusion of pertinent demographic variables, rather than a list of demographic variables. Certain demographic variables may be pertinent to the measurement of the construct which are not included in the current list. Moreover, certain constructs may have much literature which demonstrates no differences between demographic subgroups. For example, crystallized IQ would likely not require additional revalidation on samples of 70 year old patients when it was validated on a group of 50

year old patients, as the measurement properties between these two groups are quite consistent between the groups. This change to the language fits in well with the current language proposed on lines 525 through 530.

Second, we have added in the term "racial", as many demographers have now distinguished racial and ethnic identity as two different considerations (including the U.S. Census Bureau).

Line 288-300, in particular 295

"PRO instrument item generation is incomplete without patient involvement."

Suggested revision:

Item generation includes establishing the content to be covered by the items, generating item wording, evaluating the completeness of item coverage, performing initial assessment of clarity and readability. PRO instrument item generation is incomplete without patient involvement.

Motivation:

A clear definition of the term *item generation* would be helpful in clarifying this section and such a definition should also be put in the glossary. The text above is an attempt to provide such a definition. The original text could be interpreted to mean that patients should generate the item wording. However, if item wording comes directly from patient statements, considerable editing is often necessary to make the wording clear and unambiguous in a questionnaire context. Thus, patient involvement in evaluating completeness of content coverage and item clarity and readability should be sufficient.

Line 298-300

"The FDA plans to review instrument development (e.g., results from patient interviews or focus groups) to determine whether adequate numbers of patients have supported the opinion that the specific items in the instrument are adequate and appropriate to measure the concept."

Suggested change:

"The FDA plans to review instrument development (e.g., results from patient interviews or focus groups) to determine whether the items cover all aspects of the concept identified by patients, that item content is acceptable and understandable to patients, and that enough patients have been included to make the results generalizable to the populations in question."

Motivation:

The original text could be misinterpreted to suggest that adequacy and appropriateness of items is supported if enough patients support the opinion that the specific items in the instrument are adequate and appropriate. This is a too simplistic description of the approach to evaluate content validity.

Line 302-308

"Items that ask patients to respond hypothetically or that give patients the opportunity to respond on the basis of their desired condition rather than on their actual condition are not recommended. For example, in assessing the concept performance of daily activities, it is more appropriate to ask whether or not the respondent performs specific activities (and if so, with how much difficulty) than whether or not he or she can perform daily activities (because patients may report they are able to perform a task even when they never do so). Of course, it would be critical to know that each item refers to something that patients actually do."

Suggested revision:

"Items should be appropriately selected to accurately assess the desired domain, which could be 'physical function/disability', 'pain', 'fatigue', 'emotional distress', or others. Time frames, response categories, and context should be appropriate for the particular domain. Patient evaluations may be validated against external observation, e.g., actual observed ability to perform a described activity. Such external validation provides a strong test of the measurement instrument and is recommended, when feasible. Use of

items which are developed from very well-validated 'legacy' instruments or the instruments themselves is encouraged, especially when the FDA has a substantial experience with these instruments and items."

Motivation:

The original wording seem to recommend against standard practice for many PRO domains such as physical functioning and general health perceptions, where people are being asked to assess their abilities or their health. However, there is substantial evidence to support the validity of such approaches. To name one example, standard scales for both physical functioning and general health perception are strong predictors of subsequent mortality – a relationship that would be unlikely, if the patients' responses were strongly biased by a tendency to report "on the basis of their desired condition rather than on their actual condition". For "performance of daily activities" (we assume that this refers to the concept also called physical functioning) the advice of the guidance is to: 1) ask whether the patient performs the specific activity, and 2) then ask about the difficulty of performing this activity. The problem with this approach is that if a patient does not perform a particular activity, we have no way of knowing whether he or she is incapable of doing this activity or whether he or she just has chosen not to do it within the chosen recall period. This would create a tremendous problem of scoring the data that would need to be solved either by sophisticated techniques to tailor the items and the scoring to each patient or by restricting the items to activities that would be performed by everyone, if they were capable of it. This latter solution would create severe ceiling problems that would substantially reduce the usefulness of the instrument. By seeking to improve something that does not seem to be a problem in practice, the guidance would in this particular case introduce more measurement problems than it would solve.

Line 339-343

"PRO instruments that require patients to rely on memory, especially if they must recall over a period of time, or to average their response over a period of time may threaten the accuracy of the PRO data. It is usually better to construct items that ask patients to

describe their current state than to ask them to compare their current state with an earlier period or to attempt to average their experiences over a period of time."

Suggested change

"For PRO instruments that require patients to rely on memory, care must be taken to select an appropriate recall period. Long recall periods for everyday symptoms may threaten the accuracy of the PRO data. It is usually better to construct items that ask patients to describe their current state than ask them to compare their current state with an earlier period."

Motivation

The implicit assumption in the original text that cognitive processing will hamper validity and accuracy ignores the fact that any response process relies on cognitive processing and to some extent on memory. Even if the item refers to the "current state", patients have to define the time period which is referred to, and those definitions will vary. The original wording is too general; it does not clarify the term *"period of time",* and can be read as a general recommendation of diaries over other standard PRO questionnaires. Both approaches have well described advantages and disadvantages.

Line 367-369

"Response options do not bias the direction of responses (e.g., offering one negative choice, one neutral choice, and two or more positive choices on a scale makes it more likely for patients to respond that they feel or function better). "

Suggested change:

"Response options do not bias the direction of responses (in a post-intervention evaluation e.g., offering one negative choice, one neutral choice, and two or more positive choices on a scale makes it more likely for patients to respond that they feel or function better). "

Motivation:

Not all response scales can be bidirectional and balanced (e.g., pain rating scales). In a design with baseline and follow-up assessment, the issue of balanced response options is not a problem since real outcome is the difference between baseline and follow-up. Thus, the issue of balanced response scales is only relevant in designs that rely solely on post-intervention assessment.

Line 397-398

"Format refers to the exact appearance of the instrument."

Suggested revision:

"Format refers to the exact appearance of the questionnaire (or survey)."

Motivation:

Because the Glossary defines an "instrument" as being a means to capture data plus information that supports its use (e.g., documentation on scoring, analysis and interpretation of results), the word "instrument" may be too broad here, since materials such as user manuals and scoring documents generally are not seen by patients . Thus, the word "questionnaire" might be substituted. However, "questionnaire" as defined in the Glossary is limited to questions "shown to a respondent", which may be too narrow in the context of this paragraph. Another term (for example, "survey") might be added to the Glossary, to describe any means (e.g., questionnaire, diary, interview script) that is used to collect PRO data.

Line 416-422

"A scoring algorithm creates a single score from multiple items. Equally weighted scores for each item are appropriate only when the responses to the items are relatively uncorrelated. Otherwise, the assignment of equal weights will overweight correlated items and underweight independent items. Even when items are uncorrelated, assigning equal weights to each item may overweight certain items if the number of response options or the values associated with response options varies by item. The same

weighting concerns apply with added complexity when combining domain scores into a single overall score."

Suggested wording:

"When a scoring algorithm is used to create a single score from multiple items, care should be taken to ensure that psychometric requirements for multi-item scoring are fulfilled. Item selection and weighting should ensure that the score adequately represent the concept of interest. Equally weighted scores for each item is a widespread approach with considerable robustness, but care should be taken to document the validity of the approach or any other approach chosen. Inclusion of items that are conceptually too similar can lead scores to be overly influenced by a single domain. Care should also be taken when creating scores from items where the number of response options or the values associated with response options varies by item. The same weighting concerns apply with added complexity when combining domain scores into a single overall score."

Motivation:

The intention of the original statement is probably to avoid the case when some domains are overly weighted in global indices by the inclusion of many similar items from the same domain. However, taken at face value the text is not correct according to psychometric research. Within item response theory, equal weighting of items is justified when items confirm to the so-called Rasch model. This is also true when items differ in their number of response choices (regardless of the sometimes heated debate on Rasch models versus other IRT models, all IRT experts would agree with the previous statements). An implication of this is that items can be combined without weighting when they have similar item discrimination and fulfill requirements of unidimensionality and local independence. Even if items are strongly correlated, an unweighted score is valid if these requirements are fulfilled. Further, a large body of research shows that equal weighting is robust to deviations from the requirement of equal item discrimination. Within classical psychometrics, equal weighting is justified when items have roughly equal variances and item-total correlations. However, items can also be highly correlated

in this case. This revision is important since the original wording would erroneously indicate that the scoring of highly valid and reliable PROs is wrong.

Table 4

"Have patients similar to those participating in the clinical trial confirmed the completeness and relevance of all items?"

Suggested revision

Does empirical evidence support that the measurement range and item content are relevant for the patient group participating in the clinical trial and that the items cover all aspects of the domain in question?

Motivation:

The original text seems to enforce particular approaches to assessing the completeness and relevance of all items (e.g., focus groups). While focus groups can be valuable in test development, the requirement that a similar group of patients should confirm the completeness and relevance of items every time a trial is launched seems counter-productive. For example, for well-validated scales with a strong conceptual model (e.g., a scale of depression), should some items be dropped if one focus group questions their relevance? This could have dramatic consequences for the ability to compare results between studies and thus spoil the ability to interpret and generalize the results.

Line 491

"*Test-retest reliability is the most important type of reliability"*

Suggested revision:

"*Test-retest reliability is the most generally applicable way of assessing reliability"*

Motivation:

There are not different types of reliability but different ways of assessing reliability. The advantage of test-retest reliability is that it can be applied to single items measures.

Line 495-496

"Internal consistency reliability, in the absence of test-retest reliability, does not generally constitute sufficient evidence of reliability for clinical trial purposes."

Suggested revision:

"Internal consistency reliability for multi-item scale, in the absence of test-retest reliability, is sufficient evidence of reliability for clinical trial purposes only if the scoring assumptions have been carefully evaluated, and the domain in question does not exhibit large day-to-day variations."

Motivation:

Available evidence shows that internal consistency reliability is a good estimator of reliability if the basic scoring assumptions are fulfilled. Domains with a considerable day-to-day variation will have less power in clinical trials (because of the larger variation in change scores) even if the instruments are reliable according to psychometric definitions of reliability. This issue pertains to the domain in question, not to the measurement instrument.

Line 520

When a concept is expected to change, the values for the PRO instrument measuring that concept should change.

Suggested revision:

When patient experience of a concept is expected to change, the values for the PRO instrument measuring that concept should change.

Motivation:

The concept being measured should not change; rather how the patient experiences the concept might change.

Line 547-548

"An MID is usually specific to the population under study."

Suggested revision:

An MID can depend on the score level and may differ between populations. Therefore, the use of an MID in a particular population with a particular score range should be justified.

Motivation:

While MID may vary by population, much evidence suggests that MID for generic measures are probably fairly population independent, once the possible variation in score range is taken into account. In the general discussion below, we present analyses to support this point. While this topic is under-researched, the original statement is too strong given current evidence.

Line 586

On the other hand, if the PRO instrument is to be used in an entirely new population of patients, a small randomized study to ascertain the measurement properties in the new population may minimize the risk that the instrument may not perform adequately in a phase 3 study.

Suggested revision:

On the other hand, if the PRO instrument is to be used in an entirely new population of patients, a sufficiently-powered study using a representative sample from the new population to ascertain the measurement properties may minimize the risk that the instrument may not perform adequately in a phase 3 study.

Motivation:

Most studies to determine the psychometric properties of an instrument are not randomized. We propose that a "representative sample" from the new population be studied, rather than a random sample. It is prudent to encourage the use of sufficient

power in psychometric studies, including revalidation in a new sample. The word "small" connotes a sample that may not be sufficient. Feldt and Charter have many publications on the minimum sample sizes needed for sufficient stability of the psychometric estimates. Otherwise, a point estimate of .85 for reliability could have a lower bound clearly in the unacceptable range. Additionally, please see our General Comments (below) on a framework for validating the invariance of two populations on an instrument.

Line 599

A single domain from a multiple domain PRO is administered without the other domains.

Suggested revision:
Suggest deleting this statement.

Motivation:
It is not unusual for one or more scales from a multiple domain PRO to be used in a clinical trial. For example, a single scale from the Sickness Impact Profile (SIP) might be included in a PRO instrument to measure a specific concept of interest, but the entire 136-item SIP would not be included in the study. If the measurement properties of a single scale from a multiple domain PRO have been shown to be adequate in the population of interest, it is not clear why an additional study of that scale alone is necessarily required.

Line 612-613

"Patients in the proposed trial have a disease, condition, or severity level that is different from that of the patient population used for instrument development and validation."

Suggested revision:
"Patients in the proposed trial have a disease or condition that is different from that of the patient population used for instrument development and validation."

Motivation:

Invariance literature has shown that severity of a condition does not have an impact on the measurement properties between the severity levels (the "less severe" and "more severe"). Moreover, as long as each item has variance, floor and ceiling effects shouldn't constrain the psychometric aspects between the severity levels. Extreme differences between the severity levels (such as mentally impaired with an IQ of 40 compared to genius level with 140 IQ) create special circumstances, but such disparities rarely apply to medical conditions (wherein a ceiling of "normal" level is achieved).

Line 752

We suggest the following additions to the guidance:

"*Data quality should be evaluated before final data analysis is conducted and may also be monitored at regular intervals during the trial. Principles for data quality analysis should be specified in the protocol. Possible components of data quality analysis include: proportion of missing responses, proportion of invalid responses, inter-item and item-total correlations, item discriminant validity and internal consistency reliability.*

Motivation:

In our collaboration with companies conducting clinical trials, we have become aware of quality problems, even for studies submitted to the FDA and to leading clinical journals. Such problems include errors in scoring of data, which can totally invalidate the results but can easily be detected by simple tests. We therefore believe that the FDA could do the field a great service by insisting on better procedures for quality control in data collection and scoring. In our general discussion below, we present some simple ideas for monitoring data quality.

Lines 1059-1060

An HRQL measure captures, at a minimum, physical, psychological (including emotional and cognitive), and social functioning.

Suggested revision:

An HRQL measure captures, at a minimum, physical, psychological and social functioning.

Motivation:

While there are some diseases in which cognitive functioning is expected to be impaired, many, if not most, people do not have noticeable decrements in cognitive functioning. Routine measurement of cognitive functioning in unimpaired populations adds unnecessary respondent burden and cost, and cognitive functioning has not been included in a number of widely-used generic measures, including the COOP charts, Duke Health Profile (17-item version), EQ-5D, FACT-G, and SF-36. The issue of PRO measurement for patients who are cognitively impaired is addressed separately within the document, in Section E.2.

Line 1067

"*Item – An individual question, statement or task that is evaluated by the patient to address a particular concept"*

Suggested revision:

"*Item – An individual question, statement or task (and its standardized response options) that is evaluated by the patient to address a particular concept"*

Motivation:

Including response options in the definition of an item is helpful in emphasizing that response choices is an integral part of the item that needs to be carefully evaluated.

General discussion

Comments on the use of Minimally Important Differences (MID)

What is a Minimum Important Difference (MID)?

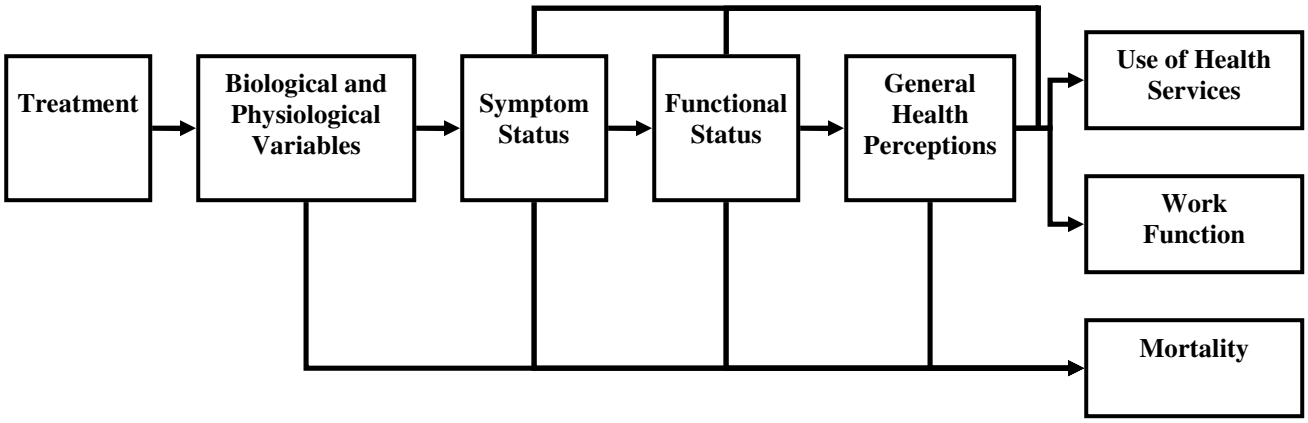
Investigators, motivated by clinicians as well as by regulatory agencies, have found the need to differentiate between an important change in a target instrument and a trivial change, to define the smallest meaningful change in a score, what is often called the "minimum important difference" or MID. A popular definition of MID is "the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's (health care) management" [3]. This definition seems linked to the experience and treatment of individual patients. The medical literature on MID recognizes that different considerations apply to the evaluation of mean group differences/score changes and to the evaluation of individual patient differences/score changes [4]. Unfortunately, this has not been reflected in the definition of MID. We therefore think the separation offered in the FDA guidance is very useful and important: "… *it is appropriate for a critical distinction to be made between the mean effect seen (and what effect might be considered important) and a change in an individual that would be considered important, perhaps leading to a definition of a responder*." (lines 540-543): The concept of MID will therefore in particular be relevant for decisions about sample size and study design, while the "responder definition" will be useful for treatment decisions in clinical practice (and to some extent for the interpretation of clinical trial results). However, both concepts should emphasize the value of the patient's own perspective and link the patient's view with that of clinicians; and both should be readily understood by clinicians and researchers[4].

Conceptual Frameworks for PRO Assessment and Deriving MID

We propose that the conceptual frameworks discussed in the guidance should also be used in evaluating MID (see an example in Figure 1). Such frameworks make important distinctions between domains of health and their operational definitions. Figure 1

portrays a specific-generic continuum [5] of PRO outcomes and links such outcomes to other variables useful for evaluating MID (also called anchor-based MID evaluation [4]). As one moves from the left to the right, the measures change from being the most highly specific biological processes, to disease-specific or generic symptoms, to generic measures of functional impact, role function and use of health services.

Figure 1. PRO Conceptual Framework (adapted from [5])

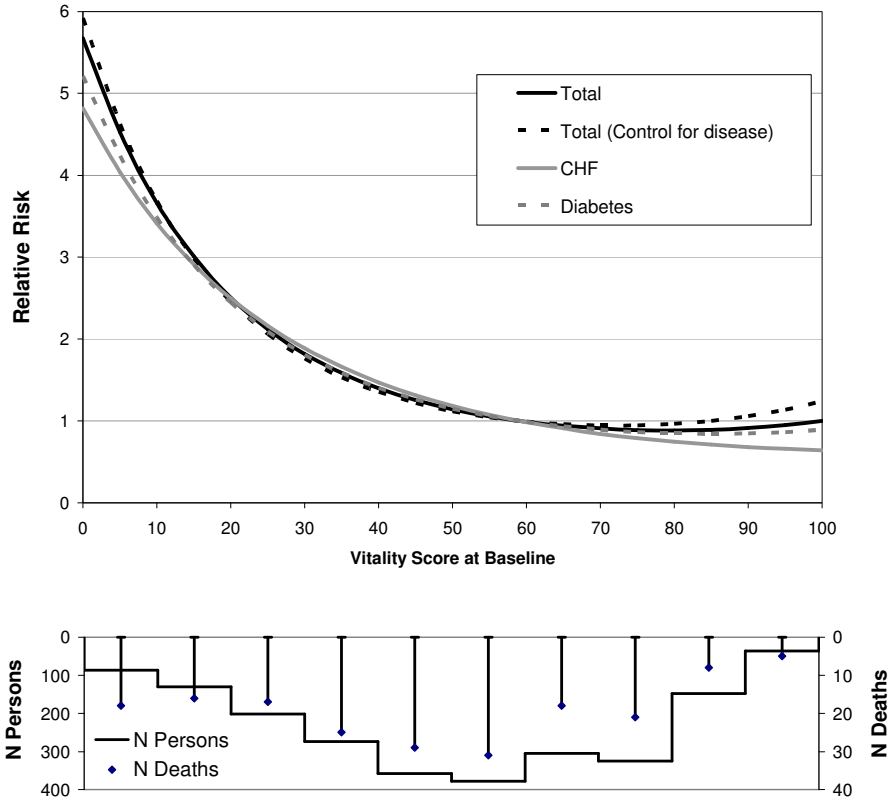


This conceptual framework also makes useful distinctions for evaluation of MID. PRO instruments that measure disease-specific impact on HRQOL may require population-specific estimates that are derived from widely accepted clinical anchors. However, an important strength of generic instruments lies in their generalizability across specific sample populations, so that investigators can make meaningful comparisons between: (1) different studies of the same disease condition; (2) different conditions; and (3) patient groups and normative benchmarks without the condition. Such comparisons, by definition, require estimates of MID that are robust with respect to sample characteristics.

A corollary issue is that a specific magnitude of change in a generic measure may have very different clinical implications depending on where a group is at baseline. For example, a change of 5 points on a physical functioning scale may have little or no real clinical implications if patients are close to normative levels at baseline. However, the same 5-point change may make the difference between being able to walk unassisted or

requiring a wheelchair if patients are starting from a lower baseline value. Thus, when presenting MIDs, the score levels for which the MID apply should be specified.

To illustrate this point, we present data from the Medical Outcomes Study, using the SF-36 Vitality (VT) score as a predictor of 7-year mortality using four different regression (proportional hazards) models: (1) total sample (n=2,199); (2) total sample with control for disease group (n=2,199); (3) CHF patients only (n=156); and (4) Diabetes patients only (n=398). All analyses controlled for gender and age group. The comparison group is a VT score of 59, the mean of the General Population (VT score distribution and deaths shown in the bottom part of the graph). All analyses show a significant effect of the Vitality score, but also a significant quadratic effect. This means that score differences above the population mean of 59 are not predictive of subsequent mortality, while score differences below 59 are highly predictive of mortality. Further, the lower the score, the more important is a score difference of a certain magnitude (Bjorner et al, in preparation).



Once the score level is taken into consideration, we do not find large variations in MID by population. In preparation of this response, we analyzed public use data from 519,035 participants in the Medicare Health Outcomes Survey (HOS). In 14 separate analyses, we performed anchor-based analyses for respondents reporting 14 different diseases to test whether the anchor-based MID would be stable across disease groups. The very large sample sizes in the HOS make this a very robust analysis. We evaluated MID for the SF-36 Physical Component Summary (PCS) using logistic regression analyses with PCS as the independent variable and 2 year mortality as the dependent variable (the anchor). Based on logistic regression results, we calculated the PCS score difference associated with a 20% increase in mortality risk (please see table below). With one exception (CHF) the results show remarkable similarity across diseases, supporting an MID of 3 points for PCS if a 20% increase in mortality risk is regarded as significant. Analyses choosing a 50% increase in mortality risk, as the threshold for minimal importance, show the same kind of stability across disease groups.

PCS score differences as predictor of mortality at 2 year follow-up. Medicare Health Outcomes Survey (N=519,035)

				MID for PCS based on increase in mortality	
Disease group	Total N	(Deaths)	Beta	20% increase	50% Increase
Depression	37618	(5499)	-0.0621	2.9	6.5
AMI	48206	(7759)	-0.0538	3.4	7.5
Angina or CAD	72153	(10033)	-0.0567	3.2	7.2
Any Cancer	59477	(9781)	-0.0567	3.2	7.2
Arthritis hand or wrist	161596	(14437)	-0.0567	3.2	7.2
Arthritis hip or knee	185084	(16438)	-0.0567	3.2	7.2
CHF	31325	(8478)	-0.0440	4.1	9.2
COPD	59733	(9200)	-0.0562	3.2	7.2
Diabetes	81439	(10033)	-0.0532	3.4	7.6

PCS score differences as predictor of mortality at 2 year follow-up. Medicare Health Outcomes Survey (N=519,035)

				MID for PCS based on increase in mortality	
Disease group	Total N	(Deaths)	Beta	20% increase	50% Increase
GI Problems	25364	(2688)	-0.0532	3.4	7.6
High BP	255713	(23675)	-0.0564	3.2	7.2
Other heart condition	97831	(12746)	-0.0585	3.1	6.9
Sciatica	109436	(9118)	-0.0572	3.2	7.1
Stroke	39049	(7453)	-0.0509	3.6	8.0

Results on MID from consensus groups of clinicians have differed between different populations, but also for different studies of the same population. The likely explanation is that the clinical consensus method is not very reliable and strongly depends on the information presented to the clinicians.

One way of handling the dependence of MID on score level is to think of MID as a function rather than a single number. This is particularly important when deriving estimates of MID for generic PRO instruments, where one may use clinically-meaningful anchors that apply to a much broader population than those typically encountered when using disease-specific instruments. An MID function associated with a measure rests on the assumption that different populations may start at different baseline values on a measure, but that there exists a consistent relationship between the measure and the anchors that is robust to sample characteristics. Early phase and/or pilot studies can be used to estimate baseline values of a measure that one might expect in a clinical trial. These can be used as the basis for calculating MID from trial to trial, but the MID function itself does not change across populations. Determining if the relationship between a PRO measure and a chosen anchor is sensitive to potential covariates (type of disease condition, age, gender, etc) is a statistical issue that is well-handled using a

number of classical methods (e.g. regression, GLM) and should be checked in many populations. Limiting an MID estimation to a single clinical population designed to reflect the sample used in a drug trial creates interpretive difficulties. The inclusion/exclusion factors (e.g., medication washout periods) that shape a clinical trial population are designed for purposes of showing efficacy (not effectiveness), and therefore may lead to erroneous estimates that do not generalize to the typical primary care setting.

Comments on Population-Specific Validation of PRO Instruments

Population definition:

It is currently difficult to understand the definition of a "population" within the guidance and what makes one group different from another. Nevertheless, the guidance make clear that differences between the PRO development population and the intended population for a clinical trial require specific validation. We believe it will prudent to provide clarification of what constitutes a substantial difference between populations. Differences between populations that have been empirically demonstrated in other research to be negligible on the outcome should not warrant a comprehensive revalidation effort. For example, prior evidence that RA and Scleroderma patients respond similarly on the HAQ-DI, along with evidence that a highly related daily performance measure has been validated on RA patients, may be ample evidence to support use of the daily performance measure on Scleroderma patients.

Invariance between samples:

We were pleased to read the statistical recommendations for using PRO in studies, and the necessary psychometric evaluations for affirming a new instrument. However, nothing was noted of the benefits of invariance testing to assure that instruments have similar psychometric properties in different populations [6;7]. Many studies assessing the similarity between two groups (including translated versions of an instrument) provide insufficient statistical accuracy. We encourage the FDA to provide details of IRT and SEM-based invariance assessment for these various procedures. A few excellent references are [8-11].

Comments on Data Quality Evaluation

The first major objective of the analysis of PRO data in a clinical trial is to provide evidence to support the scaling and interpretation of scores from the PRO instrument. Many PRO instruments consist of multi-item scales that are scored using the method of summated ratings developed by Likert [12]. The method of summated rating scales has been widely adopted for the scoring of PRO measures because of its simplicity and success in yielding reliable scores. For this reason we will focus on simple methods for evaluating summated rating scales, based on classical psychometrics. Sophisticated methods for evaluating measurement assumptions and data quality have also been developed using item response theory (within a health context see e.g.. [13;14]). To compute a score using the summated rating method, scores assigned to responses are simply summed (in some situations after recoding items so that all items in a scale are in the same direction). However, this simplicity is based on a number of assumptions that must be tested. These tests determine the appropriateness of including an item in a particular scale and whether it is appropriate to simply sum item scores to estimate and interpret the scale score. Furthermore, these tests lend support to the validity of the measurement model developed for the PRO instrument and provide a means to evaluate whether errors were made in the coding, processing and scoring of the PRO scales.

The first step in evaluating a summated ratings scale is to determine the extent of missing and out of range data. A scale score cannot be estimated with confidence if there is a large amount of missing data. A large amount of missing data for a particular item or items may indicate a problem with the wording of the item(s), with the wording of the response choices, or that respondents simply did not understand what was being asked with the question. Item means and standard deviations should also be examined. Under traditional Likert scaling criteria, item means and standard deviations should be roughly equivalent within a scale. However, depending on the purpose of measurement, the item means may be expected to be non-equivalent. For example, in measuring a wide range of physical activities, from self care to strenuous activities, item means will generally vary, because most populations will differ in their underlying ability to perform such activities.

The multi-trait item-scale correlation matrix allows for a number of assumptions traditionally associated with Likert scaling to be examined including item internal consistency and item discriminant validity. The first assumption is that an item should be substantially linearly related to the underlying concept being measured by the scale (test of item internal consistency). This assumption is tested by examining the correlation between an item and its scale, after correcting for overlap [15]. Item internal consistency is considered substantial and satisfactory if the item correlates 0.40 or more with its hypothesized scale (after correcting for overlap). A less conservative standard of 0.30 may be appropriate for evidence of item internal consistency for instruments in development. Items should correlate positively with their hypothesized scale otherwise there may be an error in the coding or handling of the item (for example, item response values requiring reverse scoring).

The second assumption, item discriminant validity, focuses on the integrity of hypothesized item groupings that are specified in the measurement model developed for a PRO instrument. It is not sufficient enough to demonstrate that an item measures what it is supposed to measure (item internal consistency), it is also important to show that an item does not measure other concepts (scales). Item discriminant validity is supported if the correlation between an item and its hypothesized scale is significantly higher than the correlations between that item and all other scales scored from the PRO instrument in a multi-trait item-scale correlation matrix.

Another property of each scale that is investigated is reliability. While reliability is discussed in the Guidance in relation to instrument development and testing, we would add that internal consistency reliability may also be used as part of the evaluation of data quality.

In summary, evidence to support the assumptions underlying the construction and scoring of PRO instruments is vital in documenting the quality of data in clinical studies and in supporting the interpretation of PRO scale scores. These simple tests outlined above should be considered as part of the analysis plan designed for clinical studies and

implemented regularly at each time point in the study to ensure that the PRO data captured meets minimum psychometric standard for scoring, analysis and interpretation.

References

1. Bjorner JB, Fayers PM, Idler EL. Self-Rated Health. In Fayers PM, Hays RD, eds. *Assessing Quality of Life*, Oxford: Oxford University Press, 2005.

2. Fayers PM,.Sprangers MA. Understanding self-rated health. *Lancet* 2002;359:187-8.

3. Jaeschke R, Singer J, Guyatt GH. Measurement of Health Status. Ascertaining the minimal clinically important difference. *Controlled Clinical Trials* 1989;10:407-15.

4. Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR. Methods to explain the clinical significance of health status measures. *Mayo Clin.Proc.* 2002;77:371-83.

5. Wilson IB,.Cleary PD. Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. *JAMA* 1995;273:59-65.

6. Groenvold M, Bjorner JB, Klee MC, Kreiner S. Test for item-bias in a quality of life measure. *J.Clin.Epidemiol.* 1995;48:805-16.

7. Bjorner JB, Kreiner S, Ware JE, Jr., Damsgaard MT, Bech P. Differential item functioning in the Danish translation of the SF-36. *J.Clin.Epidemiol.* 1998;51:1189-202.

8. Meade AW,.Lautenschlager GJ. A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods* 2004;7:361-88.

9. Reise SP, Widaman KF, Pugh RH. Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychol.Bull.* 1993;114:552-66.

10. Van de Vijver F,.Hambleton RK. Translating tests: Some practical guidelines. *European Psychologist* 1996;1:89-99.

11. Vandenberg RJ,.Lance CE. A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods* 2000;3:4-70.

12. Likert R. A technique for the measurement of attitudes. *Archives of Psychology* 1932;140:5-55.

13. Bjorner JB, Kosinski M, Ware JE, Jr. Calibration of an item pool for assessing the burden of headaches: an application of item response theory to the headache impact test (HIT). *Qual Life Res* 2003;12:913-33.

14. Bjorner JB, Kosinski M, Ware JE, Jr. The feasibility of applying item response theory to measures of migraine impact: a re-analysis of three clinical studies. *Qual Life Res* 2003;12:887-902.

15. Howard KI,.Forehand GG. A method for correcting item-total correlations for the effect of relevant item inclusion. *Educ Psychol Measur* 1962;22:731-5.