Published in final edited form as:

Med Care. 2019 May ; 57(Suppl 5 1): S38–S45. doi:10.1097/MLR.0000000000001111.

Can Methods Developed for Interpreting Group-Level Patient-Reported Outcome Data be Applied to Individual Patient Management?

Prof Madeleine T. King,

QOL Office, Level 6 North, Lifehouse (C39Z), University of Sydney, Sydney NSW 2006, Australia;
Tel: +61(0)434164438; Fax +61 (0)290365292; madeleine.king@sydney.edu.au

Amylou C. Dueck, PhD, and

Mayo Clinic, 13400 E. Shea Blvd., Johnson Res Bldg – Biostatistics, Scottsdale, AZ 85259 United States, dueck@mayo.edu, Dueck.amylou@mayo.edu

Dennis A. Revicki, PhD

Patient Centered Outcomes, Evidera, 7101 Wisconsin Ave., Suite 1400, Bethesda, MD 20814 USA, Telephone: (301) 654-9729, Fax: (301) 654-9864, Dennis.revicki@evidera.com

Abstract

Background: Patient-reported outcome (PRO) data may be used at two levels: to evaluate impacts of disease and treatment aggregated across individuals (group-level) and to screen/monitor individual patients to inform their management (individual-level). For PRO data to be useful at either level, we need to understand their clinical relevance.

Purpose: To provide clarity on whether and how methods historically developed to interpret group-based PRO research results might be applied in clinical settings to enable PRO data from individual patients to inform their clinical management and decision-making.

Methods: We first differentiate PRO-based decision-making required at group versus individual levels. We then summarise established group-based approaches to interpretation (anchor-based and distribution based), and more recent methods that draw on item calibrations and qualitative research methods. We then assess the applicability of these methods to individual patient data and individual-level decision-making.

Findings: Group-based methods provide a range of thresholds that are useful in clinical care: some provide screening thresholds for patients who need additional clinical assessment and/or intervention, some provide thresholds for classifying an individual's level of severity of symptoms or problems with function, and others provide thresholds for meaningful change when monitoring symptoms and functioning over time during or after interventions. Availability of established cut-points for screening and symptom severity, and normative/reference values, may play into choice of PRO measures for use in clinical care. Translatability of thresholds for meaningful change is

Correspondence to: Madeleine T. King.

This paper is part of the PRO-cision Medicine Methods Toolkit paper series funded by Genentech.

The PRO-cision Medicine Methods Toolkit paper series was presented during a symposium at the 2018 Annual Conference of the International Society for Quality of Life Research (Dublin, Ireland)."

more problematic because of the greater reliability needed at the individual-level versus group-level, but group-based methods may provide lower bound estimates. Caution is needed to set thresholds above bounds of measurement error to avoid “false positive changes” triggering unwarranted alerts and action in clinic.

Conclusions: While there are some challenges in applying available methods for interpreting group-based PRO results to individual patient data and clinical care - including myriad contextual factors that may influence an individual patient’s management and decision-making - they provide a useful starting point, and should be used pragmatically.

Keywords

patient-reported outcomes; minimal important difference; minimal important change; responder definition; interpretation guidelines

INTRODUCTION

Turning patient-reported outcome (PRO) data into useable information is a critical step in processing PRO data - if PRO results are not interpretable, they are of little use. PRO data are useful at two levels: group-level and individual-level. Group-level PRO data are used in research (e.g. comparative effectiveness research, clinical trials) and quality assurance settings. Methods for PRO assessment and their interpretation have evolved in these contexts over the past 30 years (1, 2). For almost as long, there has been interest in the use of PROs in clinical practice to improve patient care (3–5). This interest has spread widely in recent years, fuelled by technological developments enabling rapid PRO data capture, processing and sharing (6). Many governments at local, state and national levels around the globe are now mandating PRO use across health care systems (7). We must now hone our ability to turn individual PRO data into usable information for individual clinical decision-making.

This paper considers how methods historically developed to interpret group-based PRO research results can be applied to PRO data from an individual patient for that patient’s clinical decision-making. We first provide some background by summarizing group-level versus individual-level applications of PROs in terms of data collection context and purpose, the decision for which PRO data are collected, and the PRO comparison implied/required. Available methods for developing interpretation guidelines, cut-points, minimal important differences and change, and responder thresholds are then described. We then critique each method’s applicability to individual-level PRO data and decision-making.

Group-level versus individual-level applications of PROs

‘Individual-level’ refers to individual patient’s PRO data, e.g., used in routine care to manage that patient. ‘Group-level’ refers to aggregated individual-level data, used in research to draw inferences from samples about populations of patients and in health service contexts to describe and compare population health outcomes. Table 1 summarizes the application of PRO measures at group and individual levels, distinguishing between applications that require interpretation of PRO scores *between* groups/individuals versus *within* groups/individuals.

Interpretation of PRO scores *between* groups is required across a broad spectrum of health research and health service evaluation settings. The question is whether one group's experience differs from another's, not only in terms of statistical significance but also clinical relevance. Groups may be defined by treatment (e.g., randomized or observational studies), presence/absence of health conditions (e.g., case-control studies, normative comparisons), sociodemographic characteristics (e.g., to assess health inequality), etc. At the group level, guidance is required about how to interpret differences in group means or medians, or proportions of individuals in various health states (e.g., clinical severity). Means/medians may be of absolute scores or change scores, and proportions may be defined as scores that fall within given ranges of the PRO scale (absolute score values, e.g. severity levels) or proportions changed by a given threshold (e.g. 'responders'). The minimally important difference (MID) between groups can also be used during study design to calculate sample sizes for research studies. These group-level purposes require guidance in what is a clinically meaningful difference in absolute scores and change scores, and cut-points that define levels of severity on PRO scales and responders in terms of clinically important change.

Analogous applications at the individual level involve comparisons of an individual relative to other individuals or groups at a point in time. Examples include screening to identify patients who are at-risk of poor health outcomes or who warrant further investigation (e.g., for definitive diagnosis of anxiety, to identify cause of symptoms such as pain or dysnea, etc.) or intervention. For these purposes, cut-points on a given PRO scale are typically used to classify an individual either as a 'case' or 'normal', and may also be used to determine a patient's level of symptom severity or presence of a health problem. Comparison of the individual's PRO score with normative data may also be used to describe the impact of disease for the individual relative to a general population (8).

Interpreting change in PRO scores over time *within* groups is required not only as described above (i.e., comparing change between groups), but also in single-group studies, e.g., survivorship studies. Here, guidance is needed to interpret mean or median change within a group. Further, as per the FDA's PRO Guidance, "Regardless of whether the primary endpoint for the clinical trial is based on individual responses to treatment or the group response, it is usually useful to display individual responses, often using an *a priori responder definition*, i.e. the individual patient PRO score change over a predetermined time period that should be interpreted as a treatment benefit." (9, p24). The minimally important change (MIC) and expected proportion of responders may also be used to calculate sample sizes for single-group study designs.

Analogous applications at the individual level involve monitoring a patient through treatment and follow-up. Interpretation issues include identifying what degree of change represents real change as opposed to measurement error, represents clinically relevant improvement or deterioration, and warrants change in clinical management (e.g., starting, continuing, or stopping a treatment).

METHODS

Available methods for interpreting group-level PRO results

Because group-level data are typically analyzed with inferential statistics, it is important to distinguish statistical significance from clinical importance. Specifically, what is the smallest difference or change in PROs that still has clinical relevance. Below this threshold, even if a difference or change is statistically significant due to large sample size, it is of no consequence clinically (e.g. it won't change clinical practice).

The minimal clinically important difference was first defined as: “*The smallest difference ... 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 management*” (2, p408). Several similar terms have emerged since, differing only slightly in definition and generally estimated and used in similar ways (10). For clarity in this paper, we use *MID* to refer to the smallest clinically meaningful *difference between groups or individuals* and *MIC* to refer to the smallest clinically meaningful *change over time within a group or an individual*.

More recently, there has been interest in identifying responder definitions: “a score change in a measure, experienced by an individual patient over a predetermined time period that has been demonstrated in the target population to have a significant treatment benefit” (9, p37). Whether the threshold for a responder should be the same or larger than the MIC **is a question deserving further attention across a range of treatments and patient populations.**

Anchor-Based Methods—Anchor-based methods use an external criterion (anchor) to determine what patients or clinicians consider to be the MID or MIC and relate changes on the PRO measurement instrument to this criterion (10–12). **Anchors that reflect degree of treatment benefit can be used to determine PRO-thresholds for responder definitions** (9, p29). Criteria for selecting anchors include relevance to the disease, clinical acceptance and validity, and evidence of relationship with PRO scores (e.g., correlation of 0.3 or more) (12). As estimates for MID/MICs tend to vary among anchors and datasets, multiple independent anchors are recommended to triangulate MID/MIC values (11, 12).

Anchors can be used to derive cut-points (see Shi et al. (13) for examples) as well as MID and MICs (see Yost & Eton and Maringwa et al. (14, 15) for examples). Anchors can also be patient-based, such as patient global ratings of severity or change, or actual changes in PRO measures that have a demonstrated MID/MIC in the target patient population.

The best anchors for estimating MICs are those that identify patients who have changed in some clinically meaningful way; e.g., patients' retrospective global rating of current disease severity or change, data about the course of health over time (e.g., disease severity, change in patients who have received a treatment of known efficacy), or clinical criteria (e.g., degree of symptom control, response to treatment, etc.). A common approach to developing MICs is to have study participants at follow-up rate how much they changed since baseline using a patient global impression of change scale to assign subjects into several groupings reflecting

no change and small and large positive and negative changes in **health status**. Patients reporting getting ‘a little better’ constitute the minimal change group (i.e., MID); ratings of ‘much better’ or ‘very much better’ might be used to define responders. For example, Osoba et al. determined interpretation guidelines for small, moderate and large changes in EORTC QLQ-C30 subscale scores based on a patient global rating of change (16). Jaeschke and colleagues (2) concluded that global rating of change provided support for an estimate for a clinically meaningful difference that could be applied to both groups and individual patients. However, there are methodological criticisms of this approach, with some evidence of recall bias (17). Guyatt et al. (18) recommend that the global ratings of change are only valid where they demonstrate correlations with both pre- and post-test PRO scores. Despite these caveats, global ratings of change remain a widely-used anchor for interpreting changes in PRO measures (18).

As explained by Shi et al. (13), various analyzes are used to derive MIDs/MICs using anchors. One class of analyzes focusses on estimating mean PRO scores of groups defined by anchors (e.g., t-tests, MANOVA, regression methods); the other approach involves receiver operating characteristic (ROC) curves. Originally applied to determine the ability of a diagnostic test to detect true cases of disease (in turn determined by a gold standard method), ROC curve approaches can be applied to PRO data to identify cut-points for a range of purposes including thresholds that assists in initial screening for additional clinical assessment/intervention, identifying patient level of symptom, or alert thresholds for use in routine symptom monitoring. In this context, the PRO measure is considered the diagnostic test, and an anchor functions as the gold standard. The anchor distinguishes persons who have less/more severe symptom experience or poor/good function (for MIDs) or are importantly improved/deteriorated from persons who are not importantly changed (for MICs) (see Shi et al for details (13)). Various cut-points on the PRO measure’s scale are used to classify patients as poor/good (MID) or improved/worsened (MIC), and the cut-point with the optimal sensitivity, specificity, or other ROC-based metric is taken as an estimate of the MID/MIC. The ROC approach has been applied in two ways: originally, the anchor was a clinical criterion (e.g., 19); increasingly, the anchor is a patient-reported global transition question (e.g., 20, 21, 22).

Distribution-Based Methods—Distribution-based methods relate the observed difference or change on a PRO scale to variability in scores among patients (sample distribution) or measurement error of the PRO scale (11). A mean difference between groups may be expressed in terms of between-person standard deviation (SD) units, while a mean change within a group may be expressed in terms of within-person SD units – these are both versions of effect size (1). The standard error of measurement (SEM) of the PRO scale has also been used to gauge the degree of change that has occurred. These approaches facilitate interpretation of the magnitude of difference or change, but not whether that difference/change is important to patients and clinicians (23).

Cohen’s general guidelines for small, moderate and large effect sizes (1) are widely used, and have been empirically confirmed for two cancer-specific health-related quality of life measures (24–26). Some researchers have suggested that the $\frac{1}{2}$ SD estimate (27) or that the SEM (28, 29) may approximate a MID/MIC for some PRO instruments. Empirical evidence

shows a tendency to converge to the $\frac{1}{2}$ SD criteria as being meaningful to patients (27, 30). While this magnitude of change is certainly clinically important, it is not necessarily minimal and may set a criterion for efficacy beyond what is achievable for a given treatment.

The use of the SEM to approximate the MIC is based on study-specific observations that one SEM approximately equalled estimated MIDs, rather than what the SEM intrinsically represents and its original purpose (23) (explained below).

Therefore, when applied to group-based contexts, distribution-based measures are recommended as supportive information for MID, MIC and responder estimates derived from a range of anchor-based methods (12, 23).

Item calibrations from IRT and Rasch analysis—The use of item response theory (IRT) and Rasch analysis methods to develop PRO measures has opened possibilities of borrowing scale judgment methods used from educational research. Bookmarking is a method commonly used to identify thresholds for educational test scores. Cook et al. (31) explain and illustrate use of bookmarking methods to derive cut-points for levels of symptom severity and functional status for PROs, and to assign labels to score ranges that communicate symptom level (e.g., mild, moderate) or level of function (e.g., no problems, mild problems).

Cook et al. developed the related Idio Scale Judgment method for estimating responder thresholds (32). Participants with multiple sclerosis were asked to compare their current fatigue level to fatigue severity described in seven short form summaries (SFSs) that were located, based on item calibrations (derived with IRT analyzes), relatively close to their current fatigue levels. The subjects judged these SFSs scenarios as being comparable, more, or less fatigue than their current fatigue level. The distances were calculated between patients' current fatigue level and the location of SFSs they endorsed as a Neuro-QoL fatigue scale change that would make a difference in daily life. This approach is promising, but requires large samples covering a wide range of symptom/function levels, and IRT analyzes and calibration parameters to identify the short form summaries.

Browne and Cano (33) illustrate a more direct approach to using item content and item calibration parameters (generated via Rasch analysis) to help women considering breast reconstruction options to interpret their own PRO scores on the BREAST-Q in relation to research results using the same PRO measure.

Post Study Exit Interviews—Surveys and one-on-one interviews with study participants after clinical trial completion can be used to directly capture the patient perspective on clinically meaningful change in PRO measurements (34, 35). The survey questions can focus on overall outcomes and/or specific PRO endpoints assessed in the clinical trial. Participants are asked whether they experienced an improvement and whether that improvement was meaningful or important to them. The observed baseline to endpoint changes in primary and key secondary endpoints can be examined and compared with those who report no improvement to gain insight into clinically meaningful changes from the patient perspective. Interviews after study completion or early discontinuation can add

additional insight into the patient experience and flesh out interpretation of the quantitative survey findings.

Koochaki et al. (36) surveyed 242 women with female sexual dysfunction completing the RECONNECT phase 3 clinical trials, and conducted individual telephone interviews with 80 of these women. The survey assessed treatment benefit overall and in specific endpoint domains (i.e., sexual desire, etc.). The semi-structured interviews focused on treatment benefit, features of the treatment, and impact of improved sexual function on partner relationships, emotional functioning and well-being. The study results further confirmed responder definitions for the primary endpoints in the clinical trials.

Major challenges to designing and conducting these studies relate to the logistics in ensuring that treatment dropouts are captured, study subjects are surveyed or interviewed within 1–2 weeks of study completion and before roll-over into any open-label treatment period, training study staff and interviewers, and arranging interviews and surveys with study subjects.

DISCUSSION

Applying available methods to individual data

We now consider how the methods described above can be applied in clinical settings to enable PRO data from an individual patient to be used for that patient's clinical management and decision-making. As Table 2 shows, most methods provide thresholds: either for screening patients who need additional clinical assessment and/or intervention, for classifying an individual's level of severity of symptoms or problems with function, or for identifying meaningful change when monitoring symptoms and functioning during or after interventions.

Thresholds for screening (e.g., anxiety, depression) and cut-points for severity levels (e.g., symptoms) can be incorporated into algorithms that process the PRO data to prepare reports for clinicians/patients for recommended actions/self-management options, as illustrated by Girgis et al. for the PROMPT-Care system (37). Thresholds for change can likewise be included in algorithms for monitoring change over time, with potentially worrying changes triggering alerts and recommended action in reports to clinicians, as illustrated by Blackford et al. in PatientViewpoint (38). Thresholds can also be conveyed in various ways on PRO graphs to facilitate interpretation for clinicians and patients (39–41). Clinicians appreciate graphic presentation of results, particularly if graphs indicate what constitutes a clinically important change (42).

Considerations for screening

An important use of PROs in routine care is to identify patients at-risk of poor outcomes, in need of further investigation or intervention. This use is well-established for PROs such as anxiety and depression, but less so for symptoms and functioning. This use requires PRO cut-points, ideally validated against gold standard methods of identification. Methods that identify PRO cut-points using ROC curve analysis with a clinical diagnosis as the gold standard are clearly applicable, and cut-points already derived from such methods for certain

PROs are readily applicable. Availability of established cut-points for screening and symptom severity may play into choice of PRO measure for use in routine clinical care.

As screening tools are typically not 100% accurate, the usual considerations about false positives and false negatives apply (43): too many false positives (false alarms) will put an undue pressure on health services and cause inefficiency, while false-negatives (missed cases) may lead to legal action being taken by those individuals affected. Shi et al. (13) discuss considerations for optimal cut-points when screening for cancer symptoms, and one of the key take-home points of Girgis et al. (37) was “When considering the thresholds for ‘caseness’ for each PRO, each center should decide on the acceptable balance between false positives and false negatives, as this will impact on center workload.”

The challenge of MIDs and MICs

Like any group-based estimate, MIDs and MICs represent the mean value of a sample of individuals, all of whom varied somewhat in their individual MID/MIC values. The question is, can this aggregate value be applied equally to all individuals?

Even group-based mean MIDs/MICs may vary by clinical context and patient population, and whether patients are improving or deteriorating (10, 12). When applied to an individual patient’s decision-making, various context-specific considerations come into play: the specific health care decision, other benefits, side-effects and costs. The way any one patient balances these considerations will depend on his/her values and preferences – helping patients achieve this is a key goal of shared-decision-making.

These considerations should not preclude pragmatic use of MIDs/MICs in automated algorithms and PRO feedback summaries showing concerning levels - specific values are needed for these applications. But during the clinical consultation, there may be scope for individualizing the MID/MIC as part of shared decision-making, in terms of how a patient values the impact on function and daily life of certain treatment options and the symptoms that may be caused or ameliorated. For multi-item scales, as noted by Browne and Cano (33), the MID approach provides description at the level of overall concepts such as ‘physical function’ or ‘quality of life’ – their proposed method enables patients to consider the impact of a treatment on specific item-level issues covered by PROs.

Measurement error and “false positive changes”

PRO measures, like all measurement instruments, are subject to measurement error. This must be taken into account, particularly when tracking over time to avoid “false positive changes” triggering unwarranted alerts and action. By “false positive changes”, we mean apparent changes in scores, due to measurement error, that don’t reflect true change.

A fundamental difference between using PROs at group-level versus individual-level is the degree of reliability needed. Research requires less measurement reliability because sample sizes provide precision in the estimation of mean PRO levels, and measurement error cancels out across the sample (as it is, by definition, random). In contrast, at the individual level, very precise measures are needed (44). Historically, PRO measures designed for use in managing individual patients in psychology were very long, often containing 100 or more

items. With the advent of health status assessment in health research, measures were shortened. Many PRO measures available now are quite short, and therefore may not be sufficiently reliable for individual patient management. McHorney and Tarlov assessed five commonly-used PRO measures designed for research and found that none were sufficiently reliable for individual-level use: “rendering tentative (at best) clinical conclusions about an individual’s observed score at a point in time or changes in an individual’s score over time.” (45, p298). **This is a key rationale for the use of computer adaptive tests (CATs) in clinic. CATs are computer-administered PRO assessments that select the most informative items (questions) given the respondent’s level of health, function or symptom (46). Relative to static questionnaires, which always ask the same set of questions, CAT assessment achieves good reliability with relative brevity, and is therefore well suited for screening and monitoring individual patients (47).**

When applying short static forms in clinical care, it is therefore important to identify the threshold above which an observed change is likely to reliably reflect true change as opposed to being an artifact of measurement error (23, 48, 49). Hays et al explain and illustrate how three statistics can be used for this purpose: the SEM, the standard error of prediction (SE_p), and the reliable change index (RCI) (49). **The comparison of statistically significant change for a group (n=54) versus an individual illustrate two important points:** changes for an individual need to be much larger than changes for a group to be statistically significant, and they are unlikely to clinically trivial.

To avoid false-positive changes triggering unwarranted alerts and action for any particular PRO measure, a confidence interval (CI) on the SEM or SE_p could be included in the clinical care algorithm. Here, a one-sided CI is appropriate because only the upper limit of likely measurement error is relevant. The degree of confidence would also need to be decided; less than 95% confidence may be acceptable if the observed change triggers further assessment/discussion with clinician. If this CI is less than the group-level MIC, the MIC would hold. But if the CI was larger than the MIC, then the threshold for clinically important change would need to set at the CI in the algorithm.

Norms and reference values

The use of norms and reference values in clinical care is covered by Jensen and Bjorner (8). Conducting population-representative surveys is complex and expensive, and normative data are available for a limited number of PRO measures and countries. Systematic differences in PRO scores across different language/culture groups may limit usefulness of normative data in clinics that service multicultural communities. Norm-based PRO measures are intrinsically country-specific, e.g., item calibrations that underpin PROMIS T-scores may differ among countries. PROs often differ systematically by age and sex, again limiting the use of normative data unless available for age and sex subgroups. Finally, as yet there are no widely accepted standard thresholds around a reference value (e.g., number of SD units) to trigger clinical actions, outside of the PROs typically used in fields such as psychology.

Conclusions

This paper clarifies the information needed to make individual-level PRO results useful in clinical decision-making. We summarized available group-based methods and how they might be applied to identify screening and severity cut-points, MIDs, MICs and responder thresholds. Methods that identify PRO cut-points using ROC curve analysis with a clinical diagnosis as the gold standard are clearly applicable, as are methods that provide cut-points for symptom severity, and normative/reference values. Availability of established cut-points for screening and symptom severity, and normative/reference values, may play into choice of PRO measures for use in routine clinical care. Translatability of thresholds for meaningful change (MIC) is more problematic because of the greater reliability needed at the individual-level versus group-level; caution is needed to set threshold above bounds of measurement error to avoid “false positive changes” triggering unwarranted alerts and action in clinic. While there are some challenges in applying methods developed to interpret group-based PRO results to the use of PROs in clinical care, they can be used pragmatically as a useful starting point.

Acknowledgments

Madeleine King, Amylou Dueck and Dennis Revicki received honorarium payments for their contributions to this paper, and Amylou Dueck and Dennis Revicki received additional honorarium payments for their roles on the PRO-cision Medicine Methods Toolkit Steering Committee. In the past three years Amylou Dueck received honoraria and travel support from Bayer, Pfizer, and Phytogine unrelated to this work. The authors have no other disclosures or conflicts of interest in the past three years to declare.

References

1. Cohen J Statistical Power Analysis for the Behavioural Sciences 2 ed. Hillsdale NJ: Lawrence Erlbaum Associates; 1988 1–567 p.
2. Jaeschke R, Singer J, Guyatt GH. Measurement of health status: Ascertaining the minimal clinically important difference. *Control Clin Trials* 1989;10:407–15. [PubMed: 2691207]
3. Espallargues M, Valderas JM, Alonso J. Provision of feedback on perceived health status to health care professionals: a systematic review of its impact. *Med Care* 2000;38(2):175–86. [PubMed: 10659691]
4. Greenhalgh J, Meadows K. The effectiveness of the use of patient-based measures of health in routine practice in improving the process and outcomes of patient care: a literature review. *J Eval Clin Pract* 1999;5(4):401–16. [PubMed: 10579704]
5. Wu AW, Cagney KA, St John PD. Health status assessment. Completing the clinical database. *J Gen Intern Med* 1997;12(4):254–5. [PubMed: 9127232]
6. Luckett T, Butow PN, King MT. Improving patient outcomes through the routine use of patient-reported data in cancer clinics: Future directions. *Psychooncology* 2009;18(11):1129–38. [PubMed: 19319920]
7. Greenhalgh J, Dalkin S, Gooding K, et al. Health Services and Delivery Research. Functionality and feedback: a realist synthesis of the collation, interpretation and utilisation of patient-reported outcome measures data to improve patient care Southampton (UK): NIHR Journals Library; 2017.
8. Jensen R, Bjorner J. Applying PRO reference values to communicate clinically relevant information at the point-of-care. *Med Care Submitted - Medical Care* (this issue).
9. FDA. Food and Drug Administration. Guidance for Industry on Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. *Fed Regist* 2009;74(235):65132–3.
10. King MT. A point of minimal important difference (MID): A critique of terminology and methods. *Expert Rev Pharmacoecon Outcomes Res* 2011;11(2):171–84. [PubMed: 21476819]

11. Guyatt GH, Osoba D, Wu AW, et al. Methods to explain the clinical significance of health status measures. *Mayo Clin Proc* 2002;77(4):371–83. [PubMed: 11936935]
12. Revicki D, Hays RD, Cella D, et al. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol* 2008;61(2):102–9. [PubMed: 18177782]
13. Shi Q, Mendoza T, Cleeland C. Interpreting patient-reported outcome scores for clinical research and practice: definition, determination and application of cut-points. *Med Care* Submitted - Medical Care (this issue).
14. Maringwa J, Quinten C, King M, et al. Minimal important differences for interpreting health-related quality of life scores from the EORTC QLQ-C30 in lung cancer patients participating in randomized controlled trials. *Support Care Cancer* [Internet]. 2010 Available from: <http://www.isoqol.org/2009conference/pdf/AbstractsForBooklet2009.pdf>.
15. Yost KJ, Eton DT. Combining distribution- and anchor-based approaches to determine minimally important differences: the FACIT experience. *Eval Health Prof* 2005;28(2):172–91. [PubMed: 15851772]
16. Osoba D, Rodrigues G, Myles J, et al. Interpreting the significance of changes in health-related quality-of-life scores. *J Clin Oncol* 1998;16(1):139–44. [PubMed: 9440735]
17. Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. *J Clin Epidemiol* 1997;50(8):869–79. [PubMed: 9291871]
18. Guyatt GH, Norman GR, Juniper EF, et al. A critical look at transition ratings. *J Clin Epidemiol* 2002;55(9):900–8. [PubMed: 12393078]
19. Deyo RA, Inui TS, Leininger J, et al. Physical and psychosocial function in rheumatoid arthritis. Clinical use of a self-administered health status instrument. *Arch Intern Med* 1982;142(5):879–82. [PubMed: 7082113]
20. de Vet HC, Ostelo RW, Terwee CB, et al. Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. *Qual Life Res* 2007;16(1):131–42. [PubMed: 17033901]
21. Farrar JT, Young JP Jr., LaMoreaux L, et al. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. *Pain* 2001;94(2):149–58. [PubMed: 11690728]
22. Kvam AK, Fayers P, Wisloff F. What changes in health-related quality of life matter to multiple myeloma patients? A prospective study. *Eur J Haematol* 2010;84(4):345–53. [PubMed: 20041946]
23. Hays RD, Farivar SS, Liu H. Approaches and recommendations for estimating minimally important differences for health-related quality of life measures. *COPD* 2005;2(1):63–7. [PubMed: 17136964]
24. Cocks K, King MT, Velikova G, et al. Evidence-based guidelines for interpreting change scores for the European Organisation for the Research and Treatment of Cancer Quality of Life Questionnaire Core 30. *Eur J Cancer* 2012;48(11):1713–21. [PubMed: 22418017]
25. Cocks K, King MT, Velikova G, et al. Evidence-based guidelines for determination of sample size and interpretation of the European organisation for the research and treatment of cancer quality of life questionnaire core 30. *J Clin Oncol* 2011;29(1):89–96. [PubMed: 21098316]
26. King MT, Stockler MR, Cella DF, et al. Meta-analysis provides evidence-based effect sizes for a cancer-specific quality-of-life questionnaire, the FACT-G. *J Clin Epidemiol* 2010;63(3):270–81. [PubMed: 19716264]
27. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Med Care* 2003;41(5):582–92. [PubMed: 12719681]
28. Wyrwich KW, Tierney WM, Wolinsky FD. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. *J Clin Epidemiol* 1999;52(9):861–73. [PubMed: 10529027]
29. Wyrwich KW, Nienaber NA, Tierney WM, et al. Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life. *Med Care* 1999;37(5):469–78. [PubMed: 10335749]

30. Osoba D The clinical value and meaning of health-related quality of life outcomes in oncology. In: Lipscomb J, Gotay C, Snyder C, editors. Outcomes assessment in cancer: Measures, methods, and applications . Cambridge: Cambridge University Press; 2005 p. 386–405.
31. Cook K, Cella D, Reeve B. Applying bookmarking methods to define meaningful threshold for levels of symptom severity and functional status for patient-reported outcome measures. *Med Care* Submitted - Medical Care (this issue).
32. Cook KF, Kallen MA, Coon CD, et al. Idio Scale Judgment: evaluation of a new method for estimating responder thresholds. *Qual Life Res* 2017;26(11):2961–71. [PubMed: 28624901]
33. Browne J, Cano S. A Rasch measurement theory approach to improve the interpretation of patient reported outcomes. *Med Care* Submitted - Medical Care (this issue).
34. Coons C. Presented at the Sixth Annual Patient-Reported Outcome Consortium Workshop; Three novel methods for establishing responder thresholds on COAs.; April 29–30; Silver Spring, MD. 2015.
35. DiBenedetti D. Clinical trial exit interviews. Presented at the Clinical Outcome Assessments: establishing and interpreting meaningful within-patient change meeting; 2017 April 4; Washington, DC Duke-Margolis Center for Health Policy.
36. Koochaki P, Revicki D, Wilson H, et al. Exit survey of women hypoactive sexual desire disorder treated with bremelanotide in the RECONNECT studies demonstrated treatment benefit. Presented at the Annual Clinical and Scientific Meeting of the American College of Obstetricians and Gynecologists 2018 April; Austin, Texas.
37. Girgis A, Durcinoska I, Arnold A, et al. Interpreting and acting on the PRO scores from the PROMPT-Care (Patient Reported Outcomes for Personalised Treatment and Care) eHealth system Submitted - Medical Care (this issue).
38. Blackford A, Wu AW, Snyder C. Interpreting and Acting on PRO Results in Clinical Practice: Lessons Learned from the PatientViewpoint System and Beyond Submitted - Medical Care.
39. Brundage MD, Smith KC, Little EA, et al. Communicating patient-reported outcome scores using graphic formats: results from a mixed-methods evaluation. *Qual Life Res* 2015;24(10):2457–72. [PubMed: 26012839]
40. Smith KC, Brundage MD, Tolbert E, et al. Engaging stakeholders to improve presentation of patient-reported outcomes data in clinical practice. *Support Care Cancer* 2016;24(10):4149–57. [PubMed: 27165054]
41. Snyder CF, Smith KC, Bantug ET, et al. What do these scores mean? Presenting patient-reported outcomes data to patients and clinicians to improve interpretability. *Cancer* 2017;123(10):1848–59. [PubMed: 28085201]
42. Boyce MB, Browne JP, Greenhalgh J. The experiences of professionals with using information from patient-reported outcome measures to improve the quality of healthcare: a systematic review of qualitative research. *BMJ Qual Saf* 2014;23(6):508–18.
43. Petticrew MP, Sowden AJ, Lister-Sharp D, et al. False-negative results in screening programmes: systematic review of impact and implications. *Health Technol Assess* 2000;4(5):1–120.
44. Moinpour CM, Donaldson GW, Davis KM, et al. The challenge of measuring intra-individual change in fatigue during cancer treatment 2017;1(2):259–71.
45. McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? *Qual Life Res* 1995;4(4):293–307. [PubMed: 7550178]
46. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century 2000;1(9 Suppl):li28–42.
47. Porter I, Goncalves-Bradley D, Ricci-Cabello I, et al. Framework and guidance for implementing patient-reported outcomes in clinical practice: evidence, challenges and opportunities 2016;1(5): 507–19.
48. de Vet HC, Terwee CB, Ostelo RW, et al. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. *Health Qual Life Outcomes* 2006;4:54. [PubMed: 16925807]
49. Hays RD, Brodsky M, Johnston MF, et al. Evaluating the statistical significance of health-related quality-of-life change in individual patients. *Eval Health Prof* 2005;28(2):160–71. [PubMed: 15851771]

KEY TAKE-HOME MESSAGES

- While there are some challenges in applying available methods for interpreting group-based PRO research results to the use of PROs in clinical care, they provide a useful starting point.
- Available methods can be used to identify PRO cut-points for screening individuals and for setting thresholds for symptom severity and functional impact.
- In translating thresholds for meaningful change from group-level to individual level, caution is needed to set thresholds above bounds of measurement error to avoid “false positive changes” triggering unwarranted alerts and action in clinic.

Table 1

Applications of patient-reported outcome measures (PROs) at group-level versus individual-level

Interpreting PRO scores BETWEEN groups/individuals	Contexts/purposes	Group-level data	Individual-level data
		To compare outcomes between groups at a point in time (e.g., in comparative effectiveness research, randomized controlled trials, health service evaluation, epidemiology, etc.) A group could be compared to normative data to describe impact of disease (at a group level) relative to a general population.	To compare outcomes of an individual (relative to other individuals or groups) at some point in time (e.g., screening to identify at-risk patients and/or need for further investigation/intervention, or to classify an individual's level of severity of a condition [symptom, mental health, physical health] or level of function) An individual could be compared to normative data to describe the impact of disease on an individual relative to a general population.
Interpreting PRO scores WITHIN groups/individuals	Decision-making based on PRO data	Is one group's experience different from that of another group? To interpret research results: Is the observed difference between groups clinically important? To determine sample size: What is the MID between groups? This is used along with specified Type I and II error rates to determine how many patients are required to detect that MID.	Is a patient at-risk and/or in need of further investigation/intervention (i.e., above/below a threshold)? To evaluate the outcome status of an individual patient at a given point in time to guide decision-making (e.g., to decide treatment, to diagnose, to prognosticate, to refer to specialist, etc.).
	PRO comparison	Difference between groups in a point estimate, e.g. group mean PRO (absolute or change score) at a salient time point (e.g., post treatment), proportion of patients with severe fatigue, or the proportion changed by a clinically important degree (e.g., the MIC).	Patient's score, relative to pre-determined thresholds ('caseness'), normative data or reference values.
	Interpretation guidance needed	How to interpret the aggregate difference in PRO mean absolute scores, change scores, or proportions.	Thresholds for 'caseness' (e.g., anxiety, depression, pain) requiring further investigation/intervention. Is patient sufficiently 'non-normal' to be of concern or warrant further investigation/intervention.
	Contexts/purposes	To compare outcomes within groups over time (e.g., cohort studies)	Monitoring a patient over time (e.g., through treatment and follow-up)
	Decision-making based on PRO data	Does a group's experience change over time? To interpret research results: Is the observed change over time clinically important? To determine sample size: What is the MIC over time within a group? This is used along with specified Type I and II error rates to determine how many patients are required to detect that MIC.	Has the patient's PRO score changed sufficiently to warrant a change in management? E.g., does the patient require intervention? Should the patient cease treatment? Does the patient require dose reduction or escalation?
	PRO comparison	Difference within a group between two time points (e.g., pre- to post-treatment mean change, mean change from end of treatment to 2 year follow-up in a survivorship study).	Patient's PRO scores tracked over time – absolute levels and change relative to baseline
	Interpretation guidance needed	How to interpret aggregate change in PRO scores.	How to interpret change in patient's PRO scores: What degree of change represents real change, as opposed to measurement error? What degree represents clinically relevant improvement or deterioration? What degree of change warrants change in management?

Abbreviations: Minimally important change (MIC), Minimally important difference (MID), patient-reported outcome (PRO)

Table 2

How methods developed to interpret aggregated patient-reported outcome results can be used in individual-level clinical care

	Screen for at-risk patient / case	Classify an individual's level of severity	Monitoring - identify if a patient has clinically important change
Anchor-based methods			
Patient anchors			
Patient-rated global rating of severity		✓ Provide thresholds for severity	
Patient-rated global rating of change			✓ Provide thresholds for degree of change as perceived by patients
Clinical anchors			
Clinical diagnosis	✓ Provide thresholds for 'caseness' and severity		
Physician-rated function, symptoms, disease severity/impact		✓ Provide thresholds for severity	
Physician-rated change in function, symptoms, disease severity/impact			✓ Provide thresholds for degree of change in symptoms/disease severity/impact
Biomarker (e.g., HCT, imaging, FEV1, disease exacerbation [e.g., asthma])		✓ Provide thresholds for severity	✓ Provide thresholds for degree of change
Distribution-based methods			
Standard deviation (SD) metric (e.g., $\frac{1}{2}$ SD)		✓ Provide thresholds based on population variability in change	✓ Provide thresholds based on population variability in change
Standard error of measurement (SEM) SEM = SD at baseline $\times$ $(1 - \text{reliability})$			✓ Indicate threshold above which real change has occurred (i.e., beyond measurement error)
Other methods			
Norms & reference values	✓ Provide thresholds for low/normal/high score	✓ Provide thresholds for low/normal/high score	
Using item calibrations from item response theory or Rasch analysis (e.g., bookmarking)		✓ Provide thresholds for severity	✓ Provide thresholds for change
Post-study exit interviews with patients			✓ Provide thresholds for change as perceived by patients

L. Alternative approaches for determining statistical significance of individual-level change include the Reliable Change Index = observed change/ 2 SEM and a confidence interval (CI) based on the standard error of prediction (SE_p), e.g. 90% CI on SE_p = $1.64 \times \text{SD at baseline} \times (1 - \text{reliability})^2$