Fondazione Smith Kline

Personalità giuridica riconosciuta (D.P.R. 917 del 9. 9. 1982)

Aiom

Associazione Italiana Oncologia Medica

PROs in Oncologia
DISPENSA BIBLIOGRAFICA

Palazzo delle Stelline, Milano **10 Ottobre 2013**

Contemporary Clinical Trials 33 (2012) 23–28

Contents lists available at SciVerse ScienceDirect

Contemporary Clinical Trials

journal homepage: www.elsevier.com/locate/conclintrial

Discussion

Capturing patients' perspectives of treatment in clinical trials/drug development

Asha Hareendran [a,*], Ari Gnanasakthy [b], Randall Winnette [a], Dennis Revicki [c]

[a] United BioSource Corporation, London, UK
[b] Novartis Pharmaceuticals Corporation, East Hanover, NJ, USA
[c] United BioSource Corporation, Bethesda, MD, USA

ARTICLE INFO

Article history:
Received 5 August 2011
Received in revised form 27 September 2011
Accepted 30 September 2011
Available online 7 October 2011

Keywords:
Outcome measure
PRO
Patient outcome
Regulatory guidance
Questionnaire
Drug development

ABSTRACT

The patient's perspective of treatment outcomes is increasingly important to consumers and providers of healthcare. Recent studies have shown that traditional clinical endpoints may not accurately reflect the patient experience with treatment. Often patients' experience of their disease and associated treatment differs from the perspective of their physicians. When implemented with a clear and effective assessment strategy, patient-reported outcome (PRO) measures can be used to collect data directly from patients in the clinical setting. These data can be applied to a range of outcomes, such as treatment efficacy, safety, and patient satisfaction. Such information is valuable at various stages of drug development and can be used to understand the patient's perspective of the treatment for evaluating the treatment benefit of new products and to engage patients to make decisions about treatment options and ultimately to support commercialization of pharmaceutical products. Recognizing the value of these data, various regulatory agencies have recently released guidelines on how to best implement these measures in clinical trials to support label claims. The purpose of this paper is to discuss the benefits of collecting PRO data for evaluating the outcomes of treatments in clinical trials, through the product life cycle.

1. Introduction

Patients are today actively participating in making decisions about their healthcare and are seeking information about the treatment options available [1]. Thus, patients' perspective of treatment outcome is increasingly important to the consumers and providers of healthcare. For example, providers in the United Kingdom (UK) National Health Service (NHS) are now required to collect data using patient-completed questionnaires to evaluate functional outcomes for four elective surgical treatments. The data collected are being used to monitor outcomes and quality of care [2].

In order to participate in shared decision-making, patients look for information about the benefits of treatments in terms of outcomes that are meaningful to them. These outcomes are usually collected using measures that enable the patient to provide feedback directly. Any report of the status of a patient's health condition that comes directly from the patient, without interpretation of the patient's response by a clinician or anyone else, has been termed a patient-reported outcome (PRO) [3,4].

A range of methodological issues need to be considered for ensuring that PRO data collected in clinical trials meet the evidence requirements of the various consumers for this

Abbreviations: ACT, Asthma Control Test; AS, Ankylosing spondylitis; ASQoL, Ankylosing Spondylitis Quality of Life; CAPS, Cryopyrin-associated periodic fever syndrome; CHC, Chronic hepatitis; COPD, Chronic obstructive pulmonary disease; EMA, European Medicines Agency; FDA, Food and Drug Administration; HRQL, Health-related quality of life; NHS, National Health Service; PRO, Patient-reported outcome; RA, Rheumatoid arthritis; RLS, Restless leg syndrome; SF-36, Short Form-36®; TPP, Target product profile; VFQ-25, Visual function questionnaire-25.

* Corresponding author at: United BioSource Corporation, 26-28 Hammersmith Grove, London, W6 7HA, UK. Tel.: +44 208 834 9581; fax: +44 208 834 9555.
E-mail address: asha.hareendran@unitedbiosource.com (A. Hareendran).

1551-7144/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.cct.2011.09.015

information. It is important to collect data about the patients' views using tools that are reliable and valid [5,6]. The US Food and Drug Administration (FDA) has recently formalized a set of evidence standards required for PRO tools being used in clinical trials to support product label claims [7]. The European Medicines Agency (EMA) has also provided recommendations for the use of health-related quality of life (HRQL) measures in the evaluation of medicinal products [8].

Within clinical trials, the patient's view of symptoms, functioning, and HRQL may be evaluated using a variety of measures. These measures may include simple questions to measure the frequency (e.g., seizure rates in epilepsy) [9] or severity (e.g., joint pain in arthritis [10]) of a specific symptom. More complex multi-dimensional questionnaires are also used to measure health status in clinical trials. These include generic tools, such as the Short Form-36® (SF-36) Health Survey [11], which can be used across various disease areas, or symptom- and disease-specific measures that evaluate concepts that are important to patients experiencing the condition of interest. For example, the Ankylosing Spondylitis (AS) Quality of Life (ASQoL) tool [12] was developed to assess the impact of pain on HRQL and is used in trials of treatments for AS [13].

Recognizing the growing role of patient access to this evidence via traditional media, podcasts, patient support group websites, and social media, pharmaceutical companies see the need to generate evidence on endpoints relevant to patients. A recent review of clinical trial protocols demonstrated an emerging trend for assessing PRO endpoints in clinical trials [14]. There are opportunities for collecting and using the patients' perspective on their disease throughout the drug development process to understand the value of a treatment [15].

The purpose of this paper is to describe the benefits of collecting PRO data for evaluating the outcomes of treatments in clinical trials through the product life cycle.

2. PRO data are often used by healthcare decision makers and consumers of pharmaceutical products

2.1. For making decisions about treatment options

Clinicians have recognized the importance of patient reports to better understand the patient's health experience. Toward that end, clinicians have begun asking their patients to complete daily symptom diaries (e.g., pain, urinary incontinence, and dyspnea). Reviewing the data from these diaries helps the clinician monitor patient outcomes and inform decisions about treatment options.

In clinical guidelines [16,17] for the management of asthma, varying levels of "control" are defined based on the evaluation of PROs. These guidelines further suggest that evaluation of treatments for asthma include evaluation of improvement in patients' reports of daytime symptoms, limitations to activities, nocturnal symptoms/awakening, and need for rescue medication.

Understanding the value of improvements in PROs in terms of their association with more tangible and longer-term outcome is also useful to make decisions about treatment options. Studies have shown that improvements in patient outcomes can predict healthcare resource use [18], 1-year survival in patients with implantable cardioverter defibrillators [19], survival in women after a myocardial infarction [20], and mortality in men with chronic obstructive pulmonary disease (COPD) [21]. Because poor compliance to treatment, especially for chronic diseases that require regular treatment, can lead to poor outcomes and potentially also increased costs to the health system, information about outcomes that predict better compliance is valuable to make decision about treatment choices. The importance of maintaining the health status of the patient with chronic hepatitis (CHC) was demonstrated using the results of the pooled analyses of PRO data from three clinical trials. The data showed that decline in health status and fatigue were significant predictors of treatment discontinuation [22].

Sometimes it is also important to show the burden of illness on patients' lives to evaluate the need for treatment of the condition. The impact of conditions like restless leg syndrome (RLS) on patients' lives is often poorly understood. Data collected using a generic health status tool, the SF-36, in patients with RLS helped illustrate that these patients had significantly lower health status than US population norms and patients with other chronic medical conditions [23], suggesting the need for treatments to relieve the symptoms of the condition.

2.2. For evaluating treatment benefit of new products

Regulatory agencies also recognize the value of PRO endpoints in clinical trials to inform decisions about the safety and efficacy of products that are submitted for approval. The US FDA and the EMA guidelines for the development of products for many disease areas suggest the evaluation of PROs to assess the products in a holistic manner and interpret the value of improvements in physiological parameters. For example, the US label claim for a product for the treatment of rheumatoid arthritis (RA) includes information on patient-reported physical function to support the value of changes in pathophysiology of the disease [24]. In fact, the primary endpoint for clinical trials in rheumatoid arthritis consists of a mix of clinician ratings, inflammation, and PROs [25].

Almost half of the EMA product development guidance (39/81) documents suggest the use of PRO tools (e.g., symptom diaries) for the evaluation of new treatments [26]. These guidelines can be found at http://www.ema.europa.eu/ema for Europe and http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm for the US.

3. Data collected using PRO tools in clinical trials provide unique information about patients' experience of treatment

Historically, clinicians and regulatory agencies preferred physiological and clinician-assessed endpoints for clinical trials. However, changes in some traditional endpoints do not often reflect outcomes that are relevant to patients. For example, qualitative studies in patients with venous leg ulcers showed that rather than clinician-assessed duration and size of the ulcer, it was pain and altered appearance of the leg that impacted HRQL [27].

Further, improvements in physiological and clinical endpoints may not always translate to improvement in the patients' disease or condition. Weak correlations have been show between patients' report of symptoms and polysomnography in sleep apnea [28], HRQL, and lung functions in pediatric

asthma [29] and patient report of the impact of fibromyalgia and exercise capacity evaluated using the six-minute walk test and oxygen saturation and HRQL in COPD [30].

Patients' experience of their disease and associated treatment often differs from the perspective of their physicians. A paper examining clinician versus patient rating of symptom severity in gastroesophageal reflux disease from four randomized clinical trials showed low rates of agreement between clinician and patient assessments (kappa: 0.17–0.53) [31].

The patient's voice is also important in drug safety reporting. Comparing data collected using a patient-reported tool to clinicians' assessment of adverse events, it was shown that patients reported adverse event symptoms earlier and more frequently than clinicians [32]. PRO data have been used to understand the value of changes in pathophysiology by translating this outcome into terms that are meaningful from the patient's perspective. For example, to show the value of sustained clinical and biochemical remission (i.e., maintaining median serum level of acute-phase protein) in clinical trials of canakinumab in patients with CAPS (cryopyrin-associated periodic fever syndrome), data collected about patient reports of symptoms and health status showed that improvements in clinical parameters were associated with mental health and pain scores [33,34]. The visual function questionnaire-25 (VFQ-25), a PRO measure, has been suggested for inclusion in clinical trials evaluating treatments for ophthalmic disorders to demonstrate the value of improvements in visual acuity, in terms of improvement in patients' perception of the impact on their quality of life [35].

The value of PROs in clinical trials is not restricted to the assessment of efficacy and safety parameters. PRO measures evaluating patients satisfaction have been used to collect data about patients' overall experience of treatment, including devices used for administering them [36–40]. Data collected using treatment satisfaction tools have been considered for the selection of doses that lead to meaningful outcomes for patients. For example, in early clinical trials evaluating a treatment for benign prostatic hypertrophy, results of a treatment satisfaction scale were used to select a lower dose to be taken forward for testing. Despite the scores for satisfaction with efficacy being higher, the total score, which took into consideration satisfaction with efficacy, tolerability, and convenience, was lower for the higher doses (which had more adverse events) than the score for lower doses [41].

While clinical trials provide good estimates of average treatment effect, heterogeneity is common (i.e., not all subjects respond in the same manner to treatment or life experiences). Analyses of PRO measures collected in trials can be used to identify patient segments who benefit the most from treatments. Stull and colleagues [42] recently provided greater insight into treatment responses to a product being developed for COPD by identifying subgroups of responders and non-responders based on the analyses of PRO scores. They identified the subgroups of responders by examining the patterns and the sources of variability in the data about disease-specific health status. Such analyses enable researchers to identify groups of patients who benefit the most from new treatments and can help with subsequent trial designs (e.g., selection criteria, power calculations for sample size estimation).

PRO measures have also been used to collect data in clinical trials to demonstrate differentiation by showing the value of specific characteristics of products compared to existing treatments. For example, a PRO measure was used to collect data in clinical trials of a product for the treatment of asthma to show how patients perceived the expedited onset of action [43]. To demonstrate the value of a less frequent dosing regimen, once-a-month treatment for osteoporosis, data were collected using PRO measures in an observational study of postmenopausal women's compliance with treatment. The data clearly demonstrated that increased treatment administration frequency was associated with poor compliance [44].

4. Data collected using PRO tools have been used to engage patients to support commercialization of pharmaceutical products

Following registration of a product, data collected using PRO measures have been used to engage patients in a discussion around HRQL. The RAliving.com site, sponsored by the manufacturer of treatment for RA, enables RA patients to complete the SF-8™ Health Survey and compare their scores with the general US population. Patients can also complete the SF-8 at various time points to monitor changes in their health status. The completion of the PRO measure is seen as an opportunity to interact with patients online and makes the information being offered more personalized and relevant [45].

The Psoriasis Symptom Monitor [39], sponsored by a pharmaceutical manufacturer of treatment for psoriasis, was developed in the format of a convenient application that could be used on a computer or SmartPhone. The tool helps patients create a detailed record of their experience of symptoms over time. Patients are able to mark a diagram with affected areas, add photographs of their symptoms, and chart overall progress of their symptoms. Patients are encouraged to use the outputs of the tool to communicate to their physicians about the changes in their plaque psoriasis over time.

An early example of the use of PROs to change the treatment paradigm was in the management of patients suffering from asthma. The Asthma Control Test (ACT) was developed as part of the commercial strategy to promote the use of a specific treatment for asthma [46]. Patients were encouraged to complete the ACT to identify episodes of uncontrolled asthma and to use the information to communicate with their physicians. Information about the ACT was widely disseminated, and the ACT is now included as one of the options to use in clinical practice as per clinical guidelines. These are two examples of how pharmaceutical manufacturers are embracing new technology to enable patients with chronic conditions to monitor their own progress using PRO measures.

The recently published Salzberg Statement has called for “patients and clinicians to work together to be co-producers of health,” and the phrase “shared decision-making” has entered the lexicon of healthcare [47]. The opportunities for patients to access information for use in shared decision-making have grown recently, providing a forum for the pharmaceutical industry to disseminate information about treatment benefits and added value of products. Today, PRO endpoints are included in clinical trials to collect evidence of benefits that are meaningful to patients, to meet regulatory requirements, and to demonstrate the value of the treatment to inform decision-making in the clinic and for allocating healthcare resources.

5. Listening to the patient's voice during drug development PRO data is valuable at various stages of drug development

Developing a PRO assessment strategy is important early in drug development. It is important to plan the PRO strategy to obtain results from early clinical trials that will help refine this strategy for pivotal trials to be submitted to support product registration or reimbursement. In addition, early in drug development, a clear understanding of the patient's experience of disease and treatments can assist to identify medical needs and target profiles for new treatments. Such experiences collected directly from patients can help to inform new targets for drug discovery and target product profiles (TPP). This information can ensure that the data collected in trials will evaluate the effects of treatment on endpoints that are meaningful to patients.

The need for PRO data for key customers to support registration as well as market access and uptake by patients and clinicians must be considered early. It is important to be aware of regulatory requirements for PRO data. In the US, with the formalization of the evidence required for assessing PRO measures, it has been noted that while claims based on data collected using these measures continue to be approved by the FDA, the proportion of new molecular entities with PRO label claims during the post-guidance period (24.1%) was found to be lower than that of the pre-guidance period (30%) [48,49]. It is also important to consider the ways that patient-reported benefits can be translated into "value propositions" that are meaningful to payers. A review of existing PRO measures, their measurement properties, and the specific customer needs for the type and quality of evidence must be considered for developing a PRO measurement strategy.

As previously discussed, the data collected using PRO measures from early trials can be examined to further understand treatment responses and potentially identify subgroups of responders and non-responders. These types of analyses can inform the TPP and plan for the design of future trials that could enhance the positioning of the product for registration and reimbursement. PRO data collected on efficacy and tolerability in early trials can also be used to inform the selection of doses that are most acceptable to patients. Evaluation of the concepts related to satisfaction with treatment in terms of efficacy, safety, and convenience can also be used to inform dose selection. PRO tools can be used to engage patients in making decisions about their health and seeking specific treatments for their conditions, as in the ACT example discussed earlier. Data collected using these PROs can also be applied to healthcare decision-making. Evidence collected using PROs can be shared with patients and clinicians for use while making choices for new treatments or for considerations for switching treatments. When payers have to allocate scarce resources, they will need to prioritize and target groups of patients that would benefit most from treatment. PRO data can be used to identify sub-groups of patients who have differential responses to treatment.

6. Challenges for collecting PRO data in clinical trials as part of drug development

It must be cautioned that there are a number of challenges related to collecting PRO data in clinical trials. Developing a good PRO data collection strategy requires careful research as well as multiple and timely discussions with internal stakeholders (e.g., clinical development, regulatory, commercial, clinical trial operations) to agree on the value of PRO data for demonstrating the value of the product. Clinical trial sponsors usually have aggressive timelines that can affect the feasibility of conducting this research; therefore, extensive internal discussions should be conducted before the PRO strategy is ready to be implemented [15].

There has also been a shift within the industry to collect data more frequently. The FDA PRO Guidance [4] document states that "items with short recall periods or items that ask patients to describe their current or recent state are usually preferable." This requirement has led to the need for more frequent collection of PRO data as well as the need for use of electronic data capture. An advantage of electronic data capture is that it can be date- and time-stamped, so that any retrospective (or forward) completion of the diary cannot occur, as it often does with paper diaries [50]. However, unlike data collected in paper, the setup cost and time for electronic data capture requires much upfront planning and resources before the start of the trial. On the brighter side, recent years have seen the influx of more robust technology that can be set up in a shorter timeframe and could also improve the quality of PRO data collection.

With increasing demands to reduce the time and cost of drug development, sponsors are increasingly looking to emerging markets such as India, China, and Latin America [51,52]. This expansion has led to inclusion of countries outside of those traditionally included in global clinical trials. To ensure that the data from PRO tools can be pooled, the tools need to be translated using methods that ensure their linguistic and cultural validity [53], which add to the timeline for preparing for the start of data collection. Sponsors may be reluctant to support the logistical aspects of PRO data collection due to possible delay to study start time. This reluctance is compounded when the PRO tools are included to collect data for endpoints that are not primary in the trial or not intended to support regulatory approved label or promotional claims.

The increasing pressures from drug development teams to get studies started often do not allow teams to develop a robust PRO data collecting strategy. Often, these "shortcuts" result in conclusions that fall below expectation of benefit. The use of inappropriate instruments and the lack of explanation for the choice of instruments in clinical trials have been raised as criticism by some authors [54,55]. Moreover, the regulatory requirements for evidence of content validity of PRO tools that will be used to support label claims were tightened with the release of the US FDA PRO guidance in 2009 [4]. In a recent review of "reasons for PRO label claims being rejected by the US FDA" [56], "lack of fit for purpose" of questionnaires used in studies was cited as one of the most common reasons.

To overcome the challenges, it is important to plan for time and resources (e.g., funds; staff with relevant clinical and outcomes-related knowledge) to construct a robust PRO assessment strategy. Regulatory requirements in key markets must be kept in mind if the data are to be used to support label or promotional claims. With thoughtful strategies and careful implementation, good-quality PRO data can

be obtained to demonstrate the value of a product from the patient's perspective. The decision of whether to include a PRO measure in a clinical trial needs to take into account the objective of data collection, the value of the information to the consumer, and the burden of the data collection in the clinical trials (e.g., burden to patient, costs, etc.). Agreement on these decisions by internal stakeholders of the drug development process is also essential to the success of the PRO data collection strategy.

7. Conclusion

The PRO harmonization group,[1] in their recommendations about "Incorporating the Patient's Perspective into Drug Development and Communication," suggested that it is important to evaluate PROs in clinical trials because they are: 1) a unique indicator of the impact of disease; 2) essential for evaluating treatment efficacy; 3) useful for interpreting clinical outcomes; and 4) a key element in treatment decision-making [57]. PRO data can be used at all stages of product development and can be used for internal decision-making and for external communication to key stakeholders about product value. PROs can provide value evidence in terms of efficacy, tolerability, and treatment satisfaction and convenience. Most of all, PROs enable companies to find out what really matters to the patients.

Acknowledgements

We would like to thank Jennifer Petrillo, PhD, for her assistance reviewing of early drafts of this manuscript.

References

[1] Baldwin M, Spong A, Doward L, Gnanasakthy A. Patient-reported outcomes, patient-reported information: from randomized controlled trials to the social web and beyond. Patient 2011;4:11–7.
[2] Patient Reported Outcome Measures (PROMS). PROMS.
[3] Doward LC, McKenna SP. Defining patient-reported outcomes. Value Health 2004;7(Suppl 1):S4–8.
[4] Food and Drug Administration (FDA). Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims; 2009.
[5] Lipscomb J, Reeve BB, Clauser SB, Abrams JS, Bruner DW, Burke LB, et al. Patient-reported outcomes assessment in cancer trials: taking stock, moving forward. J Clin Oncol 2007;25:5133–40.
[6] McKenna SP. Measuring patient-reported outcomes: moving beyond misplaced common sense to hard science. BMC Med 2011;9:86.
[7] Burke LB, Kennedy DL, Miskala PH, Papadopoulos EJ, Trentacosti AM. The use of patient-reported outcome measures in the evaluation of medical products for regulatory approval. Clin Pharmacol Ther 2008;84:281–3.
[8] European Medicines Agency (EMA) Committee for Medicinal Products for Human Use (CHMP). Reflection paper on the regulatory guidance for the use of Health Related Quality of Life (HRQL) measures in the evaluation of medicinal products; 2005.
[9] Elger CE, Stefan H, Mann A, Narurkar M, Sun Y, Perdomo C. A 24-week multicenter, randomized, double-blind, parallel-group, dose-ranging study of rufinamide in adults and adolescents with inadequately controlled partial seizures. Epilepsy Res 2010;88:255–63.
[10] Dworkin RH, Turk DC, Peirce-Sandner S, Baron R, Bellamy N, Burke LB, et al. Research design considerations for confirmatory chronic pain clinical trials: IMMPACT recommendations. Pain 2010;149:177–93.
[11] QualityMetric. QM Bibliography. 2011.
[12] Doward LC, Spoorenberg A, Cook SA, Whalley D, Helliwell PS, Kay LJ, et al. Development of the ASQoL: a quality of life instrument specific to ankylosing spondylitis. Ann Rheum Dis 2003;62:20–6.
[13] Song IH, Heldmann F, Rudwaleit M, Listing J, Appel H, Braun J, et al. Different response to rituximab in tumor necrosis factor blocker-naive patients with active ankylosing spondylitis and in patients in whom tumor necrosis factor blockers have failed: a twenty-four-week clinical trial. Arthritis Rheum 2010;62:1290–7.
[14] Getz KA, Wenger J, Campo RA, Seguine ES, Kaitin KI. Assessing the impact of protocol design changes on clinical trial performance. Am J Ther 2008;15:450–7.
[15] Doward LC, Gnanasakthy A, Baker MG. Patient reported outcomes: looking beyond the label claim. Health Qual Life Outcomes 2010;8:89.
[16] Global Initiative for Asthma (GINA). Pocket Guide for Asthma Management and Prevention – A Pocket Guide for Physicians and Nurses: Medical Communications Resources, Inc.; 2008.
[17] National Asthma Education and Prevention Program. National Asthma Education and Prevention Program Expert Panel Report 3 (EPR-3). Guidelines for the Diagnosis and Management of Asthma Summary Report; 2007.
[18] Levenson JL, Hamer RM, Rossiter LF. Psychopathology and pain in medical in-patients predict resource use during hospitalization but not rehospitalization. J Psychosom Res 1992;36:585–92.
[19] Kao CW, Friedmann E, Thomas SA. Quality of life predicts one-year survival in patients with implantable cardioverter defibrillators. Qual Life Res 2010;19:307–15.
[20] Norekval TM, Fridlund B, Rokne B, Segadal L, Wentzel-Larsen T, Nordrehaug JE. Patient-reported outcomes as predictors of 10-year survival in women after acute myocardial infarction. Health Qual Life Outcomes 2010;8:140.
[21] Domingo-Salvany A, Lamarca R, Ferrer M, Garcia-Aymerich J, Alonso J, Felez M, et al. Health-related quality of life and mortality in male patients with chronic obstructive pulmonary disease. Am J Respir Crit Care Med 2002;166:680–5.
[22] Fine RN, Becker Y, De Geest S, Eisen H, Ettenger R, Evans R, et al. Nonadherence consensus conference summary report. Am J Transplant 2009;9:35–41.
[23] Allen RP, Walters AS, Montplaisir J, Hening W, Myers A, Bell TJ, et al. Restless legs syndrome prevalence and impact: REST general population study. Arch Intern Med 2005;165:1286–92.
[24] Abbott Laboratories. Humira. 2011.
[25] Felson DT, Anderson JJ, Boers M, Bombardier C, Furst D, Goldsmith C, et al. American College of Rheumatology. Preliminary definition of improvement in rheumatoid arthritis. Arthritis Rheum 1995;38:727–35.
[26] Mordin M, Lewis SE, Gnanasakthy A, Demuro-Mercon C, Copley-Merriman K, Fehnel S. Patient-reported outcomes as mentioned in product development guidance. ISPOR 15th Annual International Meeting; 2010.
[27] Hareendran A, Bradbury A, Budd J, Geroulakos G, Hobbs R, Kenkre J, et al. Measuring the impact of venous leg ulcers on quality of life. J Wound Care 2005;14:53–7.
[28] Weaver EM, Kapur V, Yueh B. Polysomnography vs self-reported measures in patients with sleep apnea. Arch Otolaryngol Head Neck Surg 2004;130:453–8.
[29] Ziora D, Madaj A, Wieckowka E, Ziora K, Kozielski K. Correlation of spirometric parameters taken at a single examination with the quality of life in children with stable asthma. J Physiol Pharmacol 2007;58 (Suppl 5):801–9.
[30] Yohannes AM, Roomi J, Waters K, Connolly MJ. Quality of life in elderly patients with COPD: measurement and predictive factors. Respir Med 1998;92:1231–6.
[31] McColl E, Junghard O, Wiklund I, Revicki DA. Assessing symptoms in gastroesophageal reflux disease: how well do clinicians' assessments agree with those of their patients? Am J Gastroenterol 2005;100:11–8.
[32] Basch E. The missing voice of patients in drug-safety reporting. N Engl J Med 2010;362:865–9.
[33] Koné-Paut I, Lachmann HJ, Kuemmerle-Deschner J. Improved health-related quality of life in patients with cryopyrin-associated periodic fever syndrome (CAPS) after treatment with canakinumab (ILARIS)—a fully human anti-IL-1 beta monoclonal antibody. American College of Radiology; 2009.
[34] Schlesinger N, De Meulemeester M, Pikhlak A, Yucel AE, Richard D, Murphy V, et al. Canakinumab relieves symptoms of acute flares and improves health-related quality of life in patients with difficult-to-treat gouty arthritis by suppressing inflammation: results of a randomized, dose-ranging study. Arthritis Res Ther 2011;13:R53.
[35] Csaky KG, Richman EA, Ferris III FL. Report from the NEI/FDA Ophthalmic Clinical Trial Design and Endpoints Symposium. Invest Ophthalmol Vis Sci 2008;49:479–89.

[1] A working group composed of members of the International Society for Quality of life Research (ISOQOL), the International Society for Pharmacoeconomics and Outcomes Research (ISPOR), the Pharmaceutical Manufacturer's Association Health Outcomes Committee (PhRMAHOC), and the European Regulatory Issues on Quality of Life Assessment (ERIQA).

[36] Bredart A, Sultan S, Regnault A. Patient satisfaction instruments for cancer clinical research or practice. Expert Rev Pharmacoecon Outcomes Res 2010;10:129–41.
[37] Secnik Boye K, Matza LS, Oglesby A, Malley K, Kim S, Hayes RP, et al. Patient-reported outcomes in a trial of exenatide and insulin glargine for the treatment of type 2 diabetes. Health Qual Life Outcomes 2006;4:80.
[38] Nordmann JP, Baudouin C, Bron A, Denis P, Rouland JF, Sellem E, et al. Xal-Ease: impact of an ocular hypotensive delivery device on ease of eyedrop administration, patient compliance, and satisfaction. Eur J Ophthalmol 2009;19:949–56.
[39] Centocor Ortho Biotech. STELARA Psoriasis Symptom Monitor.
[40] Keininger D, Coteur G. Assessment of self-injection experience in patients with rheumatoid arthritis: psychometric validation of the Self-Injection Assessment Questionnaire (SIAQ). Health Qual Life Outcomes 2011;9:2.
[41] Hareendran A, Abraham L. Using a treatment satisfaction measure in an early trial to inform the evaluation of a new treatment for benign prostatic hyperplasia. Value Health 2005;8(Suppl 1):S35–40.
[42] Stull DE, Wiklund I, Gale R, Capkun-Nigglid G, Houghton K, Jones P. Application of latent growth and growth mixture modeling to identify and characterize differential responders to treatment for COPD. Contemporary Clinical Trials 2011;32(6):818–28. [Epub 2011 Jul 6].
[43] Leidy NK, Mathias SD, Parasuraman BM, Patrick DL, Pathak D. Development and validation of an onset of effect questionnaire for patients with asthma. Allergy Asthma Proc 2008;29:590–9.
[44] Huas D, Debiais F, Blotman F, Cortet B, Mercier F, Rousseaux C, et al. Compliance and treatment satisfaction of post menopausal women treated for osteoporosis. Compliance with osteoporosis treatment. BMC Womens Health 2010;10:26.
[45] Bristol-Myers Squibb. A Case Study: Bristol-Myers Squibb's Use of the SF-8™ Health Survey for Rheumatoid Arthritis Screening, Awareness, and Education. Quality Metric.
[46] Nathan RA, Sorkness CA, Kosinski M, Schatz M, Li JT, Marcus P, et al. Development of the asthma control test: a survey for assessing asthma control. J Allergy Clin Immunol 2004;113:59–65.
[47] Gulland A. Welcome to the century of the patient. BMJ 2011;342: d2057.
[48] DeMuro C, Clark M, Mordin M, Evans E, Copley-Merriman K, Fehnel SE, et al. PHP96 reasons for rejection of PRO label claims: an analysis based on a review of PRO use among new molecular entities and biologic license applications 2006–2010. Value Health 2011;14:A29.
[49] Mordin M, Clark M, DeMuro C, Evans E, Copley-Merriman K, Fehnel S, et al. PHP97 PRO label claims: an analysis based on a review of PROs among new molecular entities and biologic license applications 2006–2010. Value Health 2011;14:A29.
[50] Byrom B. Innovative ePRO: tapping into the potential. Appl Clin Trials 2006;15(6):64–75.
[51] Torralba KD, Khan NA, Quismorio FP. Clinical trials and public trust: the geographical shift to the Asia-Pacific region. Int J Rheum Dis 2009;12:186–91.
[52] Vincent JL. Logistics of large international trials: the good, the bad, and the ugly. Crit Care Med 2009;37:S75–9.
[53] Wild D, Grove A, Martin M, Eremenco S, McElroy S, Verjee-Lorenz A, et al. Principles of Good Practice for the Translation and Cultural Adaptation Process for Patient-Reported Outcomes (PRO) Measures: report of the ISPOR Task Force for Translation and Cultural Adaptation. Value Health 2005;8:94–104.
[54] Guyatt GH, Veldhuyzen Van Zanten SJ, Feeny DH, Patrick DL. Measuring quality of life in clinical trials: a taxonomy and review. CMAJ 1989;140: 1441–8.
[55] Lee CW, Chi KN. The standard of reporting of health-related quality of life in clinical cancer trials. J Clin Epidemiol 2000;53:451–8.
[56] DeMuro C, Clark M, Mordin M, Evans E, Copley-Merriman K, Fehnel SE, et al. Reasons for rejection of pro label claims: an analysis based on a review of pro use among new molecular entities and biologic license applications 2006–2010; 2011. Poster #PHP67. ISPOR. Baltimore, MD2011.
57 Acquadro C, Berzon R, Dubois D, Leidy NK, Marquis P, Revicki D, et al. Incorporating the patient's perspective into drug development and communication: an ad hoc task force report of the Patient-Reported Outcomes (PRO) Harmonization Group meeting at the Food and Drug Administration, February 16, 2001. Value Health 2003;6:522–31.

EXPERT REVIEWS

A point of minimal important difference (MID): a critique of terminology and methods

Expert Rev. Pharmacoeconomics Outcomes Res. 11(2), 171–184 (2011)

Madeleine T King

Psycho-oncology Co-operative Research Group (PoCoG), School of Psychology, Brennan MacCallum Building (A18), University of Sydney, NSW 2006, Australia
Tel.: +61 290 366 114
Fax: +61 290 365 292
madeleine.king@sydney.edu.au

The minimal important difference (MID) is a phrase with instant appeal in a field struggling to interpret health-related quality of life and other patient-reported outcomes. The terminology can be confusing, with several terms differing only slightly in definition (e.g., minimal clinically important difference, clinically important difference, minimally detectable difference, the subjectively significant difference), and others that seem similar despite having quite different meanings (minimally detectable difference versus minimum detectable change). Often, nuances of definition are of little consequence in the way that these quantities are estimated and used. Four methods are commonly employed to estimate MIDs: patient rating of change (global transition items); clinical anchors; standard error of measurement; and effect size. These are described and critiqued in this article. There is no universal MID, despite the appeal of the notion. Indeed, for a particular patient-reported outcome instrument or scale, the MID is not an immutable characteristic, but may vary by population and context. At both the group and individual level, the MID may depend on the clinical context and decision at hand, the baseline from which the patient starts, and whether they are improving or deteriorating. Specific estimates of MIDs should therefore not be overinterpreted. For a given health-related quality-of-life scale, all available MID estimates (and their confidence intervals) should be considered, amalgamated into general guidelines and applied judiciously to any particular clinical or research context.

Keywords: clinical significance • health-related quality of life • HRQOL • interpretation • MCID • MID • minimal clinically important difference • minimal important difference • patient-reported outcome • PRO

The 'minimal important difference' (MID) is a little phrase with big appeal in a field struggling to interpret health-related quality of life (HRQOL) and other patient-reported outcomes (PROs). It is a deceptively simple term; a nuanced understanding of terminology and methods is needed to avoid oversimplification and misuse as the phrase gains popularity in a field looking for a simple solution to a complex problem.

This article critiques the terminology and methods of the MID, providing a historical context for the various 'how to'-focused papers, which summarize methods and provide recommendations [1–4]. It is presented in six sections, addressing this series of questions: how are various MID-related terms defined and what is their historical sequence? What is the MID used for? Why are HRQOL results difficult to interpret? How is the MID usually determined? How does the MID differ from the smallest statistically detectable difference, and how does it link clinical importance with statistical significance, sample size and power? It concludes by speculating on future directions for the MID in the field of HRQOL and PRO research and practice. The articles selected are not based on a systematic search, but on the author's personal experience, reading and a literature search that grew organically from that.

Evolution of definitions & terminology

Table 1 summarizes the evolution of MID-related definitions and terminology. In 1987, Guyatt *et al.* proposed the minimal clinically important difference (MCID) as the appropriate benchmark of important change against which to assess the responsiveness of an instrument or scale [5]. They did not define the MCID, and acknowledged the difficulty of quantifying it, suggesting that the change induced by an intervention of known efficacy could provide an initial estimate [5]. A total of 2 years later, in perhaps the most influential paper in MID history, the MCID was defined by Jaeschke,

 10.1586/ERP.11.9 ISSN 1473-7167

Table 1. Evolution of key terms and definitions related to the minimal important difference, methods used to operationalize or quantify them, and key distinctions between them.

Study (year)	Term	Abbreviation	Definition	Method used and/or key distinctions	Ref.
Guyatt *et al.* (1987)	Minimal clinically important difference	MCID	MCID not defined, but used definition of responsiveness: 'the ability of evaluative instruments to detect minimal clinically important differences'	Change induced by an intervention of known efficacy	[5]
Jaeschke *et al.* (1989)	Minimal clinically important difference	MCID	The smallest difference that patients perceive as beneficial and that would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management	Global transition item ('how much has your <domain of HRQOL> changed in the past <time period> '), with the threshold based on the change in HRQOL (measured prospectively) in patients who report minimal change (on the global transition item), either for better or for worse	[6]
Osoba *et al.* (1998)	Subjectively significant difference	SSD	The smallest change, either beneficial or deleterious, that is perceptible (discernable) to the subject	As per Jaeschke *et al.* [6], the important distinction is in the definition: meaningfulness is based entirely on the patient's self-assessment of the magnitude of change (note that 'perceptible (discernable)' is similar to the 'detectable' from Normal *et al.* [8]	[9]
Guyatt *et al.* (2002)	Minimal important difference	MID	The smallest difference in score in the domain of interest that patients perceive as important, either beneficial or harmful, and that would lead the clinician to consider a change in the patient's management	Methodology is not strictly prescribed; authors suggest corroboration across 'anchor- and distribution-based' methods. Authors note that the MID is the threshold between trivial and small-but-important change. Authors also note that 'subjectively significant' is a conceptually congruent alternative label for 'minimally important'	[1]
Schünemann *et al.* (2005)	Minimal important difference	MID	The smallest difference in score in the outcome of interest that informed patients or informed proxies perceive as important, either beneficial or harmful, and that would lead the patient or clinician to consider a change in the patient's management	Methodology is not strictly prescribed, but should be patient-based if possible (and while not specified, the definition implies those patients should be 'fully informed'). If proxies must be used, they should be instructed to focus on what they believe patients consider important (similarly, proxies should be 'fully informed')	[11]
Sloan *et al.* (2002)	Clinical significance		Goes beyond statistical significance to identify whether the statistically significant difference is large enough to have implications for patient care	Anchor- and distribution-based methods as described by Guyatt *et al.* [1] (the methods paper from the Clinical Significance Consensus Meeting Group of the Symposium on the Clinical Significance of Quality-of-Life Measures in Cancer Patients, Mayo Clinic [Rochester, MN, USA])	[66]

HRQOL: Health-related quality of life; R: Reliability of scale; SD: Standard deviation; SEM: Standard error of measurement.

Table 1. Evolution of key terms and definitions related to the minimal important difference, methods used to operationalize or quantify them, and key distinctions between them.

Study (year)	Term	Abbreviation	Definition	Method used and/or key distinctions	Ref.
Norman *et al.* (2003)	Clinically important differences	CID	Differences that are clinically important (as determined by the method of quantification), but not necessarily in any sense minimal	Anchor-based method involving longitudinal follow-up to determine whether subgroups can be identified that have clinically different outcomes, such as rehospitalization, relapse of cancer, Medical Research Council grading or different interventions	[8]
Wyrwich *et al.* (2005)	Clinically significant change		A difference score that is large enough to have an implication for the patient's treatment or care; sometimes corresponds to what a patient might recognize as a MID	Anchor- and distribution-based methods as described by Guyatt *et al.* [1]	[4]
De Vet *et al.* (2006)	Minimally important change	MIC	A change that patients would consider important to reach in their situation, dependent on baseline values or severity of disease, on the type of intervention, and on the duration of the follow-up period	Anchor-based methods are preferred, as they include a definition of what is minimally important	[67]
Norman *et al.* (2003)	Minimally detectable difference	MDD	As per Jaeschke *et al.* [6] – same definition, different term	As per Jaeschke *et al.* [6]. The important distinction is in the terminology: 'clinically important' is dropped in favor of 'detectable' to more accurately reflect the quantification method (i.e., patients who report minimal change on the global transition item)	[8]
Wyrwich *et al.* (1999)	Standard error of measurement	SEM	The standard error in an observed score that obscures the true score	$SEM = SD\sqrt{(1-r)}$ where SD = standard deviation of the sample and R = reliability of the scale A theoretically fixed psychometric property of an instrument or scale Takes into consideration the possibility that some of the observed change may be due to random measurement error	[38]
Beaton *et al.* (2001) and De Vet *et al.* (2006)	Minimum detectable change	MDC	Minimum change (at an individual level) detectable given the measurement error of the instrument (or scale)	$MDC(95\%\,confidence\,level) = 1.96 \times \sqrt{2} \times SEM$ where SEM as above, 1.96 derives from the 95% confidence interval of no change and √2 is included because two measurements are involved in measuring change (e.g., before and after an intervention or clinically significant event)	[14,15]

HRQOL: Health-related quality of life; R: Reliability of scale; SD: Standard deviation; SEM: Standard error of measurement.

Table 1. Evolution of key terms and definitions related to the minimal important difference, methods used to operationalize or quantify them, and key distinctions between them.

Study (year)	Term	Abbreviation	Definition	Method used and/or key distinctions	Ref.
Beckerman *et al.* (2001)	Smallest real difference	SRD	The smallest measurement change, that can be interpreted as a real difference (i.e., beyond zero), considering chance variation or measurement error	$SRD = 1.96 \times \sqrt{2} \times SEM$ (= MDC above)	[68]
Angst *et al.* (2001)	Smallest statistically detectable difference	SDD	The smallest mean change over time (within a group) which is statistically significantly different from zero	For a given sample size of n (number of patients for whom change is measured), two-sided type I error rate (α) and power (1-β, where β = one-sided type II error rate): $SDD = SD(z_\alpha + z_\beta)/\sqrt{(n/2)}$ where z_α and z_β are the values of the standard normal distribution (mean = 0, SD = 1) for α and β, respectively	[29]

HRQOL: Health-related quality of life; R: Reliability of scale; SD: Standard deviation; SEM: Standard error of measurement.

Singer and Guyatt 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" [6]. This definition planted the MCID firmly in a shared decision-making context. In 1993, in one of the most widely cited papers on HRQOL interpretation, Lydick and Epstein commended Jaeschke *et al.* on their 'wonderful' definition of the MCID, but noted that they "do not directly suggest an operational method for defining clinical meaningfulness" [7]. Jaeschke *et al.* did, however, allude to the fact that "clinicians who gain experience with a questionnaire develop a sense of the importance of changes seen in their patients' scores". In 2003, Norman *et al.* further noted that: "Nowhere in the operationalization of the MID approach is there a consideration of importance, or of the tradeoff between benefit and side effects or costs … Thus, the criterion may be more appropriately thought of as a minimally detectable difference (MDD)" [8]. Nevertheless, Jaeschke *et al.*'s method has become the standard for determining what people now typically call the MID.

In 1998, Osoba *et al.* used Jaeschke's global transition method, and coined the term subjectively significant difference (SSD), emphasizing the patient's self-assessment of the magnitude of change [9]. They defined the SSD as "the smallest change, either beneficial or deleterious, that is perceptible (discernable) to the subject". Although this term is not widely used, the paper is widely cited in cancer research as the basis of 10-point rule of thumb for the MID for the scales of the EORTC's core HRQOL questionnaire, QLQ-C30 [10].

In 2002, Guyatt *et al.* offered a nuanced definition of the MCID, rebranding it the MID: "the smallest difference in score in the domain of interest that patients perceive as important, either beneficial or harmful, and which would lead the clinician to consider a change in the patient's management" [1]. Even though the 'C' for 'clinical' was removed, it is still implicit in this definition. A total of 3 years later, Schünemann and Gyuatt added further qualifications: that patients should be informed; that proxy assessment may be used; and that proxies should also be informed [11].

In a summary of the findings of the 2002 Symposium on Clinical Significance of Quality of Life Measures in Cancer Patients, Wyrwich *et al.* described a clinically significant change in quality of life (QOL) as "a difference score that is large enough to have an implication for the patient's treatment or care" [4]. They noted that it sometimes might correspond to what a patient might recognize as a MID in HRQOL/PRO scores, but that such a change may not lead to a change in treatment or care regimen if it represented an improvement that should not be interrupted or a decline that cannot be prevented by other reasonable alternatives, particularly in advanced-stage disease where palliation is often the focus of cancer treatment [4]. As explored further in the 'Clinical anchors' section later, clinical anchors can be used to determine clinically important differences (CIDs) in HRQOL and PROs, but the extent to which these are MCIDs will depend on the anchor selected, how adjacent groups are defined within that anchor, and the strength of the relationship (conceptually and empirically) between the anchor and the target HRQOL domain.

The MID is becoming the dominant term in this literature, and the usage of this term demonstrates that the differences in terminology, definitions and methods are of little real consequence. For example, Ringash *et al.* reported MIDs from two studies that used exactly the same methods but whose stated aims differed: one was "to estimate the magnitude of difference in QOL that is noticeable to patients" [12], while the other was to "determine what magnitude of change in a patient-reported outcome score is clinically meaningful" [13]. Clearly there is a looseness in the

usage of MID-related terms that readers need to be aware of; Table 1 helps to make explicit the differences and similarities in terms, definitions and methods.

'Minimally detectable': in what sense?

A confusing aspect of the MID-related definitions and terminology arises from the fact that a difference in HRQOL may be 'minimally detectable' in two senses. One is owing to the limits of perception and relates to the use of a global transition item. If we consider only those patients who felt they had got better or worse by the smallest possible increment, then this is the MDD in the sense used by Norman *et al.* [8] (as explained previously). Indeed, Norman *et al.* explain their finding of an apparently universal MID being equivalent to an effect size of 0.5 standard deviation (SD) by reference to psychophysiological evidence that the human limit of cognitive discrimination is approximately one part in seven, which in many empirical circumstances is very close to half a SD. This sense is also captured in Osoba *et al.*'s term SSD, which also relies on the global transition anchor method.

The other sense of 'minimally detectable' is in terms of the limits of measurement – that imposed by measurement error, which relates to the standard error of measurement (SEM), as its name implies. It is in this sense that Beaton *et al.* [14] and de Vet *et al.* [15] use the term 'minimum detectable change' (MDC): the amount of individual-level change that must be observed before it is considered above the bounds of measurement error. In other words, it is the threshold at and above which the individual change observed on a particular scale (with fixed SEM) reflects real change in the underlying (latent) domain of interest. The importance of appreciating the distinction between the MDC and the MID is put succinctly by de Vet *et al.* [16]: "to judge whether the minimally detectable change of a measurement instrument is sufficiently small to detect minimally important changes". Since the SEM is a theoretically fixed psychometric property of an instrument or scale, then so is the MDC (as a function of SEM; Table 1).

Group versus individual differences

An important but sometimes obscure distinction in the MID literature is that of group-level differences versus within-individual changes. This distinction is explored further in the following section. In MID-related literature, the term 'MDD' is typically used in relation to the former, while 'MDC' is used in relation to the latter. If the MDC is larger than the MCID, then the measure is insufficiently precise for individual monitoring. This is of increasing relevance with emerging interest in using PROs and HRQOL scores in monitoring and managing individual patients [17]. However, this issue is immaterial at the group level, where required levels of precision for mean differences (reflected in the standard error of the mean) are provided by adequately powered sample sizes (whether mean differences between groups or mean change within groups). This issue is considered further in the section entitled 'Methods used to determine the MID'.

Uses of the MID: decision-making at the individual & group levels

Health-related quality of life and PRO questionnaires have the potential to play a key role in bringing the patient's voice to evidence-based healthcare. However, to realize this potential, we need to be able to interpret the relevance of PROs in making decisions about treatment. Such decisions are made at both the individual level, when a patient (or their clinician, acting as their agent) chooses among treatment options or decides to cease or reduce treatment, and at the group level, when clinical research is conducted to test the relative effectiveness of treatments, often testing a promising new treatment against current best practice. At both of these levels, we need to know how much of a difference in PRO or HRQOL scores matters. The difficulty is working out to whom it should matter and in what sense it should matter.

Use of the MID in shared decision-making

At the individual level, when managing and monitoring patients in routine care, we need to know how much change in HRQOL is sufficient to warrant a change in treatment, whether starting a new treatment, continuing or stopping a current treatment, or increasing or decreasing the dose. Clearly this will vary across treatment contexts, and will often involve balancing benefits against side effects, inconvenience, financial costs to the patient and other less concrete costs. In the case of a treatment aimed at slowing the progression of Parkinson's disease, if deterioration in mobility is too rapid on the current dose, the decision may be made to increase the dosage, despite side effects. In the case of palliative radiotherapy for bone metastases, it may be that an improvement in pain is needed before treatment is ceased. In the case of adjuvant chemotherapy, it may involve the trade-off of likely survival gains against the HRQOL consequences of the toxicity burden. In most cases, each aspect of benefit, harm or cost will have a threshold beyond which the treatment decision will tip one way or the other. This may be a complex decision, involving the balancing the benefits and downsides of the treatment. How these are balanced will differ from patient to patient. Ideally, each patient's decision will be made at the point that best matches that patient's preferences.

Use of the MID in research

The decision context at the group level is quite different. Typically, a randomized trial is conducted to determine the relative efficacy of two treatment options, with patients randomized to treatment. Such trials provide robust evidence to guide policy at the health service provision level and practice at the individual patient management level. How is the final decision about which is the best treatment made in clinical trials? Analogous to individual-level decisions, it will involve an often complex balancing of benefits and harms. But at the group level, we use hypothesis testing and statistical analysis; the section of this article entitled 'Statistical significance, sample size, power and the smallest statistically significant difference' explores how the MID relates to these. The question still remains: what are the appropriate thresholds to tip a decision one way or another? It seems reasonable to use the average of individual patient's thresholds (MIDs).

Another use of the MIDs at the group level is in responder analysis, and similarly, the presentation of results in terms of proportion of patients that have improved, remained stable or deteriorated. This has been recommended by various influential authors as a means to present group-based results in a way that is more meaningful to clinicians [18–20]. Yost *et al.* have suggested that the upper end of a MID range be used in such cases to account for the higher level of measurement error for an individual change score [21].

The problem of interpreting HRQOL & PRO scores

Across these various uses of the MID, the underlying challenge is meaningful interpretation of patient-reported scores, in particular, the threshold that represents the smallest difference or change that tips a particular treatment decision one way or the other. A recent review found that HRQOL results were rarely interpreted in terms of clinical significance, even in randomized controlled trials reported to a high standard [10]. So while the clinical trials community has accepted the validity and feasibility of HRQOL and PRO assessment, and while such end points are increasingly used in clinical trials, a troubling lack of competence in making sense of the results still persists.

Why is it so hard to interpret HRQOL & PRO data?

Many complex factors confound our understanding of HRQOL and PROs. First, HRQOL is intrinsically a subjective phenomenon, a perception, reliant on self-report. It may mean different things to different people, and therefore defies definition (which is partly why we have moved to the less problematic term 'PRO'). Second, a particular individual's perception of their HRQOL may vary over time as their circumstance and perspective changes. Indeed, the capacity to adapt psychologically to loss of health is a boon to the individual, but the consequent 'response shift' in their self-report of their health and QOL [22,23] is one of the great challenges to interpreting changes in HRQOL data. Third, HRQOL is an umbrella term that covers a wide range of health-related phenomena (or 'constructs'), including physical, social and emotional functioning and a variety of symptoms of disease and side effects of treatment. These are measured by a vast number of questionnaires (or 'instruments'), each of which may contain several domain-specific scales. Each scale is intrinsically different because it includes a unique set of questions (or 'items'). All these scales are somewhat arbitrary in terms of their numeric values because there is no absolute zero or standard scalar increment for phenomena such as pain, fatigue and social function. Fourth, the response options on HRQOL items are ordinal. A fairly typical set of options is: 1 = not at all; 2 = a little; 3 = quite a bit; 4 = very much. These numbers do not have interval properties, that is, the difference between 'not at all' and 'quite a bit' may not be the same as the difference between 'a little' and 'very much', which may not be twice as big as the difference between 'not at all' and 'a little'. This is true for all such self-report scales, including numbered scales anchored by two phrases such as 'none at all' and 'worst imaginable'. Fifth, there can be differences among individuals in the way they use these response scales. For example, a person who has a low threshold for pain or fatigue may rate their level as 'quite a bit', while a more stoic individual may rate the same level as 'a little' pain, and a person's pain threshold may increase with their experience of chronic pain. Sixth, questions are often aggregated into multi-item scales, and the scores from individuals are aggregated into group-level results, often presented as mean HRQOL scores. Each step away from the content of the particular questions in the scale represents a further abstraction. Finally, few people have the requisite understanding of psychometrics or the hands-on experience with HRQOL and PRO assessment methods to confidently interpret the results that arise from specific scales, and there are surprisingly few interpretation manuals available. The generic short-form health survey (SF-36) and the cancer-specific Functional Assessment of Cancer Therapy (FACT-G) provide rare exceptions [24,25].

It is therefore surprising that any sense can be made of HRQOL and PRO data. Yet, despite the odds, when well-developed and validated HRQOL and PRO questionnaires are used, remarkably sensible patterns are apparent in the resultant data. For example, when the QLQ-C30 was used to measure the HRQOL of cancer patients, those with more advanced disease typically reported more symptoms and a poorer QOL across a range of functional domains compared with those with less advanced disease [26], and when the SF-36 was used in a large population sample, the group of people who developed a new long-term health condition on average reported a decline in all but one domain of HRQOL [27]. These patterns inspired confidence that HRQOL data could be interpreted meaningfully. In 1993, Lydick and Epstein provided an insightful review and influential taxonomy of methods for interpreting QOL results [7]. A subset of these now persist as methods used to determine the MID. The following section describes these, and gives some historical perspective on each one.

Methods used to determine the MID

This section describes the four methods that are historically and currently the most commonly used methods for determining the MID, and includes some other less widely used methods. The first and fourth of these described are typically called 'anchor-based' and the second and third are called 'distribution-based', after Lydick and Epstein [7], or are alternatively termed 'externally-referenced' and 'internally-referenced', respectively [28].

Global transition questions

Patient retrospective rating of change using a global transition question (as the 'anchor' or 'external reference') was first reported in 1989 [6], and has become the most commonly used method for determining the MID. The PRO is assessed prospectively at two time points, at the second of which the subject is also asked to think back to the first time point and judge the degree of change in that particular outcome, using a single item that has a series of graded options, often this five-point version: 'much worse', 'a little worse', 'the same/ no change', 'a little better' and 'much better'. For multidomain HRQOL instruments, this is typically performed by domain. So, for example, if the MID for an emotional functioning scale is to be determined, the global change question would ask about the degree of change in emotional

functioning since the previous HRQOL assessment time point, and this would be linked with the prospectively measured change in emotional function.

Typically, the mean change of the groups that differ by 'a little' is taken as the estimate of the MID. Some authors estimate separate MIDs for improvement and deterioration, and adjust the slightly better/worse results by subtracting the mean change that occurs in the 'no-change' group, for example Angst *et al.* [29]. The latter correction, similar to the adjacent category mean difference method of Cella *et al.* [30] and Maringwa *et al.* [31], is not universally accepted. While Hays *et al.* support the practice of comparing the change in HRQOL for individuals that have been deemed to change by a minimal amount with the change observed for those who are deemed to have stayed the same (not changed), they do not support the subtraction of the latter from the former [2]. Rather, they recommend that if the mean change for the no-change group is similar to that of the minimally changed group, then the MID estimate is suspect. However, if the MID change exceeds that of the no-change group, the MID estimate is useful and does not need to be adjusted by the HRQOL change observed in the no-change group. For example, if the minimally important change group is found to have an average change in HRQOL of four points versus two points for the no-change group, then the four points is the estimated MID and two points is not enough to constitute a MID.

Note that a change deemed to be 'a little better/worse' is not explicitly important or significant in any sense, which is why Osoba *et al.* called it the SSD [9]. Such thresholds are certainly relevant to the communication between patient and clinician because they represent the degree of change where patients begin to notice an improvement or decline; clearly anything smaller cannot be relevant to the therapeutic encounter. This method has some limitations. First, because judgements are retrospective, they may be prone to response shift and recall bias [32]. Second, patients' retrospective estimates of change are more highly correlated with their present state than with their change in health state [27,33]; this has been confirmed in cognitive interviews [34]. Third, their validity as measures of change has not been formally evaluated, and fourth, they are single items and so are more prone to measurement error than multi-item scales are. Fifth, when transition scales contain more than the five possible options already listed, the cut-points used to define the MID group are somewhat arbitrary, and it is often assumed that change related to an improvement is the same as that for a decline. For example, four change groups were defined in a 15-point transition scale: trivial (- 1, 0 or 1), minimal (2, 3 or -2, -3), moderate (4, 5 or -4, -5) and large (6, 7 or -6, -7) by Metz *et al.* [35].

Finally, two points should be noted about the results of this method. First, they demonstrate that a lot of variation exists among individuals, as illustrated in figure 1 from Osoba *et al.* [9] and figure 2 of Knox and King [27]. So while the means of the various change groups generally follow the expected trend (largest mean deterioration in HRQOL in the group, which felt very much worse, through to the largest mean improvement in HRQOL in the group, which felt very much better), in each group there are likely to be at least some individuals whose prospectively measured change scores contradict their retrospectively assessed global change, and this may be a significant proportion of patients in the smallest change groups. It is unclear the extent to which this reflects the truth of how these individuals felt versus measurement error and other limitations of this method previously described.

A related issue is that sample sizes in each change group are often quite small (as the total sample size is divided into five or seven change groups), so corresponding mean change scores tend to have large confidence intervals. For example, several of the 95% confidence intervals on the mean change for the two smallest change groups in Osoba *et al.*'s figure 1 include zero. Despite this, the ballpark message from Osoba *et al.*'s article, which now echoes throughout the literature [10], is that a ten-point change is the MID, regardless of clinical context or HRQOL domain. This demonstrates that simple messages resonate more readily in the research literature than complex ones, and thereby become embedded in research practice.

Individual variation is expected in all biological phenomena, and the use of a mean MID in clinical research to calculate sample size and interpret aggregate results is consistent with practice for objective health outcomes. However, when using HRQOL to monitor and manage individual patients, the subjective and multidimensional nature of HRQOL and the personalized trade-offs that patients make mandate that patient's opinions and preferences should be sought and considered if decision-making is to be truly shared with the clinician.

Increasingly, global transition questions are being used in another way to determine MIDs via receiver-operator curve (ROC) analysis, as described in the section on 'Other less commonly used methods'. De Vet *et al.* provide an example, illustrating how the ROC approach can be used to address the question: 'How sure we are that this MID value holds for every patient?' [36].

Standard error of measurement

Another commonly used approximation for the MID is the SEM, a theoretically fixed psychometric property of a HRQOL or PRO scale. Conceptually, the SEM is a measure of the spread of observed scores of a notional individual around their true score, had that patient been repeatedly assessed on the same measurement scale, with no memory and/or response effects and while having the same underlying HRQOL or other target PRO. The estimation of the SEM does not involve a patient or proxy's input about whether a change is minimally important in any sense; it has no 'anchor' or 'external-reference point'. Thus the SEM is not really a method for estimating the MID; it is merely a convenient proxy for the MID, easily calculated from available data or published estimates of between-person SD and scale reliability (r):

$$SEM = SD\sqrt{(1-r)}$$

While some argue that test–retest reliability should be used [37], others make the case that Cronbach's α is suitable for HRQOL-related phenomena, particularly those that are highly fluid even over short time periods [38].

Historically, the SEM was used in the field of 'individual differences' in psychology. A confidence interval was constructed around an individual's observed score using the standard normal-based approximation of 68% confidence within 1 SEM and 95% confidence within SEM (or more accurately 1.96 SEM). This indicated the limits beyond which an observed change was likely to reliably reflect true change, as opposed to being an artefact of measurement error. It could also be used to determine whether an individual whose observed score fell close to a cut-point really fell above or below it. This approach is illustrated for HRQOL measures by McHorney and Tarlov in assessing whether five commonly used instruments are sufficiently reliable for monitoring and managing individual patients [37].

The SEM is now used in the HRQOL field as a convenient criterion for estimating MIDs, following validation of this approach by Wyrwich and colleagues for various measures [38–40]. These validation analyses were based on a small set of studies that compared the SEM with established MID thresholds; three studies suggested that the SEM was about the same size as the MID, while the other three suggested the MID was more than twice as large as the SEM. Wyrwich reconciled these apparent differences by considering the extent of change considered to be minimally clinically important; in the former three studies it was change ratings of 'a little better' or 'somewhat better', while in one of the latter three studies (the only one that used a global transition item as the anchor), the MCID was based on patients who felt 'a good deal better' to 'a very great deal better'. Wyrwich concluded that, in both cases, one SEM was equivalent to the change experienced by patients who felt 'a little/somewhat better' [39]. The growing number of articles reporting the SEM alongside other MID estimates provides the opportunity to further assess the generalizability of this relationship. For example, Turner *et al.* recently reported that one SEM provided a reasonable approximation to anchor-based estimates of the MID for two respiratory questionnaires [41].

Effect size

The effect size (ES) is the most general approach to MID determination and, like the SEM, it has no external reference point or anchor for interpretation. It is a 'signal-to-noise ratio': the mean difference (or change) in HRQOL divided by the variability among individuals (SD). Two ES summary statistics are commonly used to estimate the MID: a fifth and a half of a SD. The convention in the MID literature is to use the between-person SD, typically at baseline, perhaps influenced by Kaziz *et al.* who recommended this in 1989 as an aid to interpretation [42]. This statistic is also called the standardized mean difference and Cohen's D [43] and, although it is just one of many variants of the very general notion of an ES, it is commonly 'the' ES in the HRQOL literature, again probably influenced by the Kaziz paper (note its title) [42]. Further confusion in terminology can arise because of the more general medical use of the term as a synonym for 'intervention effect' or 'effect estimate'.

Historically, the ES provided a solution to the problem of interpretation across numerous scales, as needed in meta-analysis, where the same PRO (such as depression or pain) is measured on different scales. The ES standardizes all scales to a common metric; because the numerator (mean difference) and denominator (SD) are both in the same measurement units (the particular HRQOL scale), their ratio is unit-less or scale-free. This allows PRO or HRQOL effects measured on different scales to be directly compared in terms of the variability among individuals, or 'standard deviation units'. However, like the SEM, it does not answer the question of whether a difference is minimally important in any sense. For example, if a new treatment that shifts the mean HRQOL by, say, half a SD, while the current best treatment shifts it by only a third of a SD, should we update policy and practice to the new intervention?

In order to address this question, Norman *et al.* conducted a systematic review of studies that computed a MID and contained sufficient information to compute an ES [8]. A total of 38 studies yielded 62 ESs with an average ES for the MID of 0.48. They consequently proposed that an ES of 0.5 be adopted as a universal standard MID. In a reanalysis of the same data, Farivar *et al.* reported a somewhat lower average ES (0.42) due to different assumptions, inclusions and exclusions, and noted the wide variation among studies (range: 0.11–2.3) [44]. So while a universal standard MID is appealing in its simplicity, opinion remains divided about the accuracy and utility of such a generalization [44–46].

Many years ago, Cohen proposed operational definitions of small, medium and large ESs for the standardized mean difference of 0.2, 0.5 and 0.8, respectively [43]. As the title of his well-known book (still widely available 41 years after its first publication) suggests, he was motivated by the prevalence of underpowered studies in the social sciences. His guidelines are now widely used in healthcare research, not only to calculate sample sizes suitably powered to test hypotheses (as he intended), but also to interpret results. Interestingly, Cohen described his guidelines as 'arbitrary conventions, recommended for use only when no better basis for estimating the effect size is available' [25,43]. While this caveat generally seems to have been overlooked, some researchers have taken up the challenge of developing evidence-based ESs. King *et al.* used an innovative method combining systematic review of published studies, expert opinion and meta-analysis to address this issue for the widely used cancer-specific HRQOL questionnaire, the FACT-G [47]. For some domain scales, the evidence-based ESs were considerably larger than Cohen's guidelines, in which case use of Cohen's guidelines would lead to overpowered studies and to over-interpretation of the clinical significance of an observed effect. For other domain scales, evidence-based ESs were considerably smaller than Cohen's thresholds; in these cases, use of Cohen's guidelines would lead to underpowered studies and inconclusive results. King *et al.*'s results also revealed variation between cross-sectional and longitudinal results, and between domains of HRQOL. Similar conclusions were reached by Cocks *et al.* who undertook a similar exercise for the EORTC's QLQ-C30 [48]. As the ES is signal-to-noise ratio, such variations may be driven by differences in both the signal detected by the scale (reflected in the mean differences) and the variability among individuals (reflected in the SD).

Relationship of SEM & ES, & limitations for MID estimation

As noted by Wyrwich *et al.* [4], there is a relationship between the SEM and the ES; the higher reliability, the lower the ES needed

to achieve a MID. For example, for a measure with reliability of 0.75, 1 SEM implies an ES of 0.50, while for a measure with higher reliability of 0.96, 1 SEM implies an ES of 0.20. However, as noted by Hays *et al.* [2], neither ESs nor SEM provide information about the size of a difference or change in a measure that is minimally important. Evidence such as that collated by Norman *et al.* [8] and Wyrwich [39] has been used to make the case that simple guidelines generalize across measures, and therefore ESs and SEMs can be used as convenient proxies of MIDs. Hays *et al.* are more circumspect in suggesting that ESs be used to explore the extent to which MID estimates are similar or vary across instruments, and recommend that anchor-based methods should be the primary method of estimating the MID [2], as do Revicki *et al.* [3]. Such anchor-based methods include global transition items and clinical anchors.

Clinical anchors

Another method used to determine the MID, designed to aid interpretation of mean HRQOL results, is to group the HRQOL scores by clinical criteria that clinicians are familiar with, called clinical anchors [7]. This is sometimes called the 'known groups approach', where 'known' is short-hand for 'the clinical status of the groups is known' [49]. Several criteria must be satisfied for this method to work [1–3]. Clinicians should be familiar with the anchor, usually because it is widely used in assessing and/or managing patients. The anchor itself should be interpretable. There should be a theoretical basis for the relationship between the anchor and the relevant HRQOL domain(s), and an empirical correlation of at least 0.30 between the anchor and those HRQOL domain(s). Anchors with these characteristics are often used during validation to test the clinical criterion validity of HRQOL and PRO measures.

A classic anchor in cancer is the ubiquitous clinician-rated Eastern Cooperative Oncology Group Performance Status (Table 2). It is used by clinicians to rate a patient's daily activities of living. It is commonly used in cancer clinical trials as an inclusion criterion and codified in practice guidelines for chemotherapy and surgery on the basis that the patient needs to be well enough to survive these treatments. It is commonly used in validation of cancer-specific HRQOL measures. King's review demonstrated that groups with a worse performance status consistently had worse physical function, role function and cognitive function and more fatigue, nausea and pain, but the emotional and social scores did not follow this pattern, confirming its usefulness as an anchor for developing interpretation guidelines for most, but not all, HRQOL domains [26]. Clinician-rated performance status has been used as an anchor to determine MIDs (and, more generally, CIDs), cross-sectionally and longitudinally (for improvement and deterioration, separately), for various cancer-specific measures in the Functional Assessment of Chronic Illness Therapy suite [50] and, more recently, for the EORTC's QLQ-C30 [31].

By their nature, anchor-based estimates of MIDs are dependent on the choice of anchor and the strength of the relationship between the specific HRQOL domain and the anchor chosen. For example, using change in hemoglobin level as an anchor, Cella *et al.* found larger differences in fatigue and anemia-focused scales than in the total FACT-G score [30]. Furthermore, as these anchors are by their very nature clinically meaningful, anchor-based HRQOL differences are likely to be more meaningful to clinicians and researchers than to patients, thus they are CIDs. The extent to which they are MCIDs depends on the anchor selected and how adjacent groups are defined within that anchor. For example, when hemoglobin level was used as the anchor for the Functional Assessment of Chronic Illness Therapy fatigue and anemia scales, and adjacent groups were defined by trichotomizing hemoglobin level as <8g/dl, 8–9.99 g/dl and 10–11 g/dl, mean differences between the adjacent groups were similarly small and, arguably, each was minimally clinically important [30]. But when performance status (PS) was used as an anchor, and groups were defined as 0, 1 and 2–3 (the latter being combined due to small sample sizes), the mean HRQOL difference between groups PS1 and PS2–3 was two-to-three-times larger than that between groups PS0 and PS1. Arguably, the latter difference was more likely to be an MCID, while the former was a CID but not an MCID. More generally, collapsing anchor categories owing to limitations in sample size (a relatively common practice) may lead to overestimation of the MID.

Other less commonly used methods

Each of the following methods is innovative and provides for the input of various stakeholders to the judgment of what is minimally important. While they provide information-rich results, they are more logistically complex and/or labor intensive than the methods already described, which may explain why they are less commonly used.

Redelmeier *et al.* developed a method for estimating the MID that requires patients to judge themselves in relation to others with the same condition (based on between-patient differences), and found it produced similar results to the global transition method previously described (based on within-patient changes) for the Chronic Respiratory Questionnaire [51]. More recently, Redelmeier collaborated with Ringash *et al.* to apply this to cancer [12,13]. This method avoids the major problems of response shift and compounded measurement error of the global transition method.

Receiver-operator curves, commonly used to determine the ability of a diagnostic test to detect true cases of disease (in turn determined by a gold-standard method), have also been used to determine MIDs. In this context, the HRQOL measure is considered the diagnostic test and a clinical anchor functions as the gold standard. The anchor distinguishes persons who are significantly improved or deteriorated from persons who have not significantly changed. Various cut-points on the HRQOL instrument's scale(s) are used to classify patients as improved or not improved, and the cut-point with the optimal ROC characteristics (sensitivity and specificity) is taken as an estimate of the MID. The ROC approach has been applied in two ways. Originally, the anchor was a clinical criterion, as illustrated by Deyo *et al.* [52]. Increasingly, the anchor is a global transition question, as illustrated by Kvam *et al.* [53] and de Vet *et al.* [16].

Expert opinion has also been used in two ways. In the first, Wyrwich and colleagues triangulated views from expert physicians, patients and the clinicians treating these patients on

Table 2. Eastern Cooperative Oncology Group performance status.

Grade	Description
0	Fully active, able to carry on all predisease performance without restriction
1	Restricted in physically strenuous activity but ambulatory and able to carry out work of a light or sedentary nature (e.g., light house work, office work)
2	Ambulatory and capable of all selfcare but unable to carry out any work activities. Up and about more than 50% of waking hours
3	Capable of only limited selfcare, confined to a bed or chair more than 50% of waking hours
4	Completely disabled. Cannot carry on any selfcare. Totally confined to a bed or chair
5	Dead

From [69], with credit to Eastern Cooperative Oncology Group, Robert Comis MD, Group Chair.

how much change in a HRQOL measure was needed for that change to be considered a trivial, small, moderate or large clinically important improvement or decline in asthma [54], heart disease [55] and chronic obstructive pulmonary disease [56]. The expert panels participated in complex and lengthy consensus processes about what constituted clinically important differences, and devised wording for global transition questions and eligibility criteria for the patient participants, who completed questionnaires (including global transition questions) and participated in interviews bimonthly for 1 year. The clinicians treating these patients completed baseline assessments on each patient's health state and then evaluated the change in each patient's condition at subsequent visits during the next year. In asthma and chronic obstructive pulmonary disease, the patient-perceived estimates were consistent with the results of previous global change-based MIDs but were notably lower than those derived from the expert panel and the managing clinicians. In heart disease, however, they found little consensus and concluded that MID estimates depended largely on the rater's perspective and the method used. The authors nevertheless felt that this approach demonstrated the value of patient and physician perspectives and the need for improved dialogue and understanding in the interpretation and use of HRQOL results.

King *et al.* used expert judgement in another way, combining it with clinical anchors and systematic review. Three clinicians with many years of experience managing cancer patients and using HRQOL outcomes in clinical research each reviewed 71 papers that reported mean scores of the FACT-G, a cancer-specific HRQOL measure. Blinded to the FACT-G results, they considered the various clinical anchors associated with FACT-G mean differences, predicted which dimensions of HRQOL would be affected and whether the effects would be trivial, small, moderate or large. These size classes were defined explicitly in terms of clinical relevance. The experts' judgments were then linked with FACT-G mean differences and inverse-variance weighted mean differences and ESs were calculated for each size class [47,57]. Cocks *et al.* applied a similar method to the QLQ-C30 [48]. In both of these studies, variations in MIDs were found across domains of HRQOL.

Statistical significance, sample size, power & the smallest statistically significant difference

Yet another angle on the MID is its relationship with statistical significance. It is often noted that the MID is informative for calculating sample sizes, as demonstrated, for example, in the study by Cocks *et al.* [48]. Fayers and Machin provide a comprehensive description of sample size calculation for various HRQOL scale types and analysis methods [58]. In simple terms, sample size calculation determines the number of patients required to allow a reasonable chance (power, the complement of the type II error rate) of detecting a predetermined difference (which may be the MID) in the outcome variable at a given level of statistical significance (type I error or false-positive rate).

The smallest (statistically) detectable difference (SDD) is the smallest difference that can be detected as statistically significantly different from zero, given nominated type I and II error rates and fixed sample size. It is a function of these quantities and the SD of scores at baseline, as described in Angst *et al.* (Table 1) [29]. The SDD may be smaller or larger than the MID. If it is larger, then the study is underpowered to detect the MID as statistically significantly different from zero; the confidence interval will include the MID and zero. In the arthritis and rheumatism literature, the SDD of candidate outcome measures is sometimes estimated and compared with the MCID (e.g., see Angst *et al.* [29]).

This is the group-level decision-making research context. Here, the MID (or equivalently MCID) is the smallest difference that will convince clinicians to change their treatment practice or that will convince policy-makers to change their practice guidelines or the treatments they make publicly available on subsidized schedules. If, at the planning stage, sample size has been calculated to detect the MID, then the study is appropriately powered to detect the MID, and when the data are finally in, the interpretation of the results will be straightforward. Problems may arise if sample size is based on other considerations, such as when HRQOL is a secondary outcome and the trial is powered on the primary end point. Overpowered HRQOL comparisons may arise in randomized trials powered for survival end points (which typically require larger samples than HRQOL end points), and in population-based surveys or cohort studies, where large sample sizes are likely. Here, the danger is that very small HRQOL differences (clinically trivial) will be statistically significant.

Thus, statistical significance can only be used to interpret HRQOL results if the sample size was determined *a priori* on the MID. Even then, the clinical significance should be considered and discussed to provide a useful interpretation of the results for readers. However, Cocks *et al.* found that of 82 cancer randomized controlled trials reporting EORTC QLQ-C30, clinical significance was only addressed in 38% of these [10]. Where clinical significance was not addressed, reliance was usually based on

statistical significance. This misuse of statistical significance is compounded in HRQOL studies, where the multidimensional nature of HRQOL leads to multiple hypothesis testing and the associated danger of false-positive findings [58].

Expert commentary & five-year view

The occurrence of 'the MID' and related terms in the HRQOL literature has approximately trebled every 5 years over the past 20 years. Many of the recent studies are determining MIDs, either for the first time for a measure or again in another clinical context, or using MIDs to interpret the clinical significance of mean differences or to determine the proportion of patients with clinically important change. This represents progress in the interpretation of HRQOL and PRO results in general.

Interestingly, the term 'MID' (or any of the related terms in Table 1) was conspicuously absent from the US FDA's final guidance for industry on PRO measures [59]. Instead, the term 'responder definition' was used, defined as 'the individual patient PRO score change over a predetermined time period that should be interpreted as a treatment benefit.' They said that it should be determined empirically, and went on to describe the four most commonly used methods for determining MIDs described in the 'Methods used to determine the MID' section, stipulating that transition questions and clinical anchors should provide the primary evidence, with ES and SEM as supportive evidence, as per recommendations for MID determination of Revicki *et al.* [3]. It is unclear why the authors avoided the term 'MID'. They were certainly talking about something that closely resembles what others might call a MID, although with the added time dimension.

Revicki *et al.*'s stance on recommended methods and the hierarchy of evidence for MIDs [3] was also taken in the consensus statement of the Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT) [60], which covered MIDs, MCIDs and CIDs. These two guidance publications, both published in 2008, reflect current consensus on methods and therefore probably indicate future trends, at least for the next 5 years: the methods described in the 'Methods' section are likely to remain the most widely used, alone and in combination. Expert opinion may also be enlisted to quantify a range of CIDs beyond the MID, following the methods such as those pioneered by Wyrwich and colleagues [54–56], and King and colleagues [47,57].

In addition, it is likely that a plethora of MID estimates will appear in the coming years. While it may be tempting to adopt a single published MID *prima facie*, we need to be mindful that it is just an estimate, as prone to sampling variation as any other, and influenced by the method used, the patient population, the clinical context and perspective [2,3]. As Ware and Keller sagely observed in 1996, interpretability is not established by a single psychometric maneuver; rather, it develops gradually as a body of evidence accumulates with repeated experience from a variety of perspectives [61]. Simple rules of thumb for interpreting HRQOL measures are appealing, but should be used judiciously. Ringash *et al.* provide a good example of the balance required. After reporting MID estimates for different domains, with 95% confidence intervals, they quite reasonably simplified these to: "One rule of thumb for interpreting a difference in QOL scores is a benchmark of about 10% of the instrument range", adding the caveat, "Patients appear to be more sensitive to favorable differences, so an improvement of 5% may be meaningful" [13]. In summarizing our results for the FACT-G, we heeded the advice of Guyatt *et al.* to avoid misleading oversimplifications [62]. As we believed that interpretation guidelines for HRQOL scales require some flexibility to accommodate different patient groups and clinical circumstances, we summarized our results for each size class and domain as probable ranges. Furthermore, the degree of variation of component estimates within size classes in our meta-analysis highlighted the limitations of individual studies for deriving general interpretation guidelines. In addition, rather than focus on the MID, we accommodated the possibility that in some circumstances, the MID may be of a moderate absolute size, while in others it may be relatively small.

As estimates of MIDs emerge from individual studies, we need to consolidate them into a growing store of knowledge. Who should do this, and how should it be done? Some widely used instruments are managed by large organizations, such as the EORTC's Quality of Life Group and Department (for the QLQ-C30 and its modules) and QualityMetric (for the SF-36 and other measures). These are the obvious entities to take on this responsibility, preferably in the form of continuously updated interpretation guidelines. Individual researchers, or consortia such as IMMPACT [60], may have the interest and means to prepare and update reviews of available evidence about MIDs, and present them with accompanying text that explains to less expert readers the suggested use and caveats of MIDs within broader interpretation guidelines that emphasize the importance of context. Different MID estimates may be graphed to visually depict the range of estimates, with informal weighting and synthesis, as illustrated by Revicki *et al.* [3], or by meta-analysis, as illustrated by King *et al.* [47,57] and Cocks *et al.* [48] While such guidelines will lack the immediate appeal of a general rule-of-thumb, they will encourage a more sophisticated practice in the interpretation of HRQOL data, as recently recommended [2,3].

Part of the sophistication that we as a research community should aspire to is the matching of MIDs to clinical contexts and treatment decisions, as emphasized by various authors [2,3,59,63,64]. In reviewing medical product development to labeling claims, the FDA will "evaluate an instrument's responder definition in the context of each specific clinical trial" [59]. Wyrwich *et al.* provide a good example of context-specific interpretation that firstly involves the determination of thresholds based on treatment satisfaction questions and then the determination of the doses of desvenlafaxine that provide the degree of symptom relief considered important by menopausal women, as defined by the satisfaction thresholds [65]. Yet the MID history demonstrates that simple messages tend to resonate and propagate through the research literature and practice. Do time-poor researchers really want to acknowledge that there is no universal MID, that 'the MID' does not exist? As de Vet *et al.* said, "A balance needs to be struck between the practicality of a single MIC [sic] value and the validity of a range of MIC [sic] values" [16]. The point of minimal important difference is indeed elusive.

Financial & competing interests disclosure

The author has no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

Key issues

- Several minimal important difference (MID)-related terms differ only slightly in definition: minimal clinically important difference, clinically important difference, minimally detectable difference and the subjectively significant difference. Others appear similar but have quite different meanings: minimally detectable difference versus minimum detectable change. Four main methods are commonly used to estimate MIDs: patient retrospective rating of change using global transition items (patient-change anchor); commonly used clinical scales or classifications (external clinical anchors or 'known groups'); standard error of measurement; and effect size. These are described and critiqued in this article.
- Definitions and methods are summarized in Table 1 of this article. Nuances of definition of the MID-related terms are rarely of any consequence in the way these methods are applied, and the results reported and used.
- There is no global MID, although an effect size of between 0.2 and 0.5 may provide a useful ballpark guideline. For a particular patient-reported outcome (PRO) instrument or scale, the MID is not an immutable characteristic, but may vary by population and context. There is considerable variation in individual-specific MIDs.
- At the group level, the MID may need to be adjusted for the clinical context and decision at hand, whether other benefits or side effects are considered in that decision, the baseline from which the patient starts (relatively well or sick), and whether the patient is improving or deteriorating.
- At the individual level, when used in shared decision-making, the MID should be adjusted to match the patient's preferences.
- Empirical estimates are known to differ with domain-specific scales and by which method is used (particularly with clinical anchors). Therefore, specific estimates of MIDs should not be overinterpreted. For a given PRO scale, all available MID estimates and ranges should be considered and applied judiciously to any particular clinical or research context.
- PRO scales commonly used in research settings may not be reliable enough ($\alpha > 0.9$) to detect MIDs at the individual patient level because they are relatively short (typically between two and five questions in each scale). The advent of computerized adaptive testing may provide a solution if/when this technology becomes widely adopted.
- In clinical research, statistical significance cannot be used to interpret health-related quality of life and PRO results unless the sample size has been based *a priori* on a specified MID. Even then, the clinical significance should be considered and discussed to provide a useful interpretation of the results for readers.
- Future directions: multiple methods should be used to determine not only MIDs but also a wider range of clinically important differences (small, moderate and large effects), with global transition questions and clinical anchors providing primary evidence, and standard error of measurement and effect size as supportive evidence. It would be helpful for researchers if available estimates of MIDs and clinically important differences were consolidated into interpretation guidelines for specific health-related quality of life and other PRO measures, with periodic updates as further evidence emerges.

References

Papers of special note have been highlighted as:
• of interest
•• of considerable interest

▶1 Guyatt GH, Osaba D, Wu AW *et al.* Methods to explain the clinical significance of health status measures. *Mayo Clin. Proc.* 77(4), 371–383 (2002).

▶2 Hays RD, Farivar SS, Liu H. Approaches and recommendations for estimating minimally important differences for health-related quality of life measures. *COPD* 2(1), 63–67 (2005).

▶3 Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J. Clin. Epidemiol.* 61(2), 102–109 (2008).

•• **Authoritative summary of recommended methods.**

▶4 Wyrwich KW, Bullinger M, Aaronson N *et al.* Estimating clinically significant differences in quality of life outcomes. *Qual. Life Res.* 14(2), 285–295 (2005).

▶5 Guyatt GH, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. *J. Chronic Dis.* 40(2), 171–178 (1987).

▶6 Jaeschke R, Singer J, Guyatt GH. Measurement of health status: ascertaining the minimal clinically important difference. *Control. Clin. Trials* 10, 407–415 (1989).

▶7 Lydick E, Epstein RS. Interpretation of quality of life changes. *Qual. Life Res.* 2, 221–226 (1993).

•• **Insightful review and influential taxonomy of methods to interpret health-related quality of life (HRQOL).**

▶8 Norman GR, Sloan JA, Wyrwich WK. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Med. Care* 41(5), 582–592 (2003).

▶9 Osoba D, Rodrigues G, Myles J, Zee B, Pater J. Interpreting the significance of changes in health-related quality-of-life scores. *J. Clin. Oncol.* 16(1), 139–144 (1998).

▶10 Cocks K, King MT, Velikova G, Fayers PM, Brown JM. Quality, interpretation and presentation of EORTC QLQ-C30 data in randomised controlled trials. *Eur. J. Cancer* 44, 1793–1798 (2008).

• **Review of standard of reporting and interpretation for HRQOL outcomes in randomized controlled trials.**

▶11 Schünemann HJ, Guyatt GH. Goodbye (M)CID! Hello MID, where do you come from? (Commentary). *Health Serv. Res.* 40(2), 593–597 (2005).

▶12 Ringash J, Bezjak A, O'Sullivan B, Redelmeier DA. Interpreting differences in quality of life: the FACT-H&N in laryngeal cancer patients. *Qual. Life Res.* 13(4), 725–733 (2004).

▶13 Ringash J, O'Sullivan B, Bezjak A, Redelmeier DA. Interpreting clinically significant changes in patient-reported outcomes. *Cancer* 110(1), 196–202 (2007).

▶14 Beaton DE, Bombardier C, Katz JN *et al.* Looking for important change/differences in studies of responsiveness. OMERACT MCID Working Group. Outcome Measures in Rheumatology. Minimal Clinically Important Difference. *J. Rheumatol.* 28(2), 400–405 (2001).

▶15 de Vet HC, Terwee CB, Ostelo RW, Beckerman H, Knol DL, Bouter LM. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. *Health Qual. Life Outcomes* 4, 54 (2006).

▶16 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.* 16(1), 131–142 (2007).

▶17 Snyder C, Aaronson N. Use of patient-reported outcomes in clinical practice. *Lancet* 374(9687), 369–370 (2009).

18 Fayers PM, Machin D. *Quality of Life: Assessment, Analysis and Interpretation (1st Edition)*. John Wiley & Sons Ltd, NY, USA (2000).

▶19 Guyatt G, Schunemann H. How can quality of life researchers make their work more useful to health workers and their patients? *Qual. Life Res.* 16, 1097–1105 (2007).

▶20 Osoba D, Bezjak A, Brundage M, Zee B, Tu D, Pater J; Quality of Life Committee of the NCIC CTG. Analysis and interpretation of health-related quality of life data from clinical trials: basic approach of the National Cancer Institute of Canada Clinical Trials Group. *Eur. J. Cancer* 41, 280–287 (2005).

▶21 Yost KJ, Cella D, Chawla A *et al.* Minimally important differences were estimated for the Functional Assessment of Cancer Therapy-Colorectal (FACT-C) instrument using a combination of distribution- and anchor-based approaches. *J. Clin. Epidemiol.* 58(12), 1241–1251 (2005).

▶22 Norman G. Hi! How are you? Response shift, implicit theories and differing epistemologies. *Qual. Life Res.* 12(3), 239–249 (2003).

23 Schwartz CE, Sprangers MA. *Adaptation to Changing Health: Response Shift in Quality-of-Life Research (1st Edition)*. American Psychological Association, Washington, DC, USA (2000).

24 Cella D. *Manual of the Functional Assessment of Chronic Illness Therapy (FACIT) Measurement System (4th Edition)*. Evanston Northwestern Healthcare & Northwestern University, IL, USA (1997).

25 Ware JE Jr. *SF-36 Health Survey: Manual and Interpretation Guide*. The Health Institute, Boston, MA, USA (1993).

▶26 King MT. The interpretation of scores from the EORTC quality of life questionnaire QLQ-C30. *Qual. Life Res.* 5, 555–567 (1996).

▶27 Knox S, King MT. Validation and calibration of the SF-36 health transition question against an external criterion of clinical change in health status. *Qual. Life Res.* 18(5), 637–645 (2009).

28 Osoba D, King M. Interpreting QOL in individuals and groups: meaningful differences. In: *Assessing Quality of Life in Clinical Trials: Methods and Practice*. Fayers P, Hays R (Eds). Oxford University Press, Oxford, UK, 243–257 (2005).

▶29 Angst F, Aeschlimann A, Stucki G. Smallest detectable and minimal clinically important differences of rehabilitation intervention with their implications for required sample sizes using WOMAC and SF-36 quality of life measurement instruments in patients with osteoarthritis of the lower extremities. *Arthritis Rheum.* 45(4), 384–391 (2001).

▶30 Cella D, Eton DT, Lai JS, Peterman AH, Merkel DE. Combining anchor and distribution-based methods to derive minimal clinically important differences on the Functional Assessment of Cancer Therapy (FACT) anemia and fatigue scales. *J. Pain Symptom Manage.* 24(6), 547–561 (2002).

• **Illustrates methods with an insightful discussion.**

31 Maringwa JT, 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* DOI: 10.1007/s00520–010–1016–1015 (2010) (Epub ahead of print).

▶32 Ross M. Relation of implicit theories to the construction of personal histories. *Psychological Rev.* 96(2), 341–357 (1989).

▶33 Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. *J. Clin. Epidemiol.* 50(8), 869–879 (1997).

▶34 Wyrwich K, Tardino V. Understanding global transition assessments. *Qual. Life Res.* 15(6), 995–1004 (2006).

▶35 Metz SM, Wyrwich KW, Babu AN, Kroenke K, Tierney WM, Wolinsky FD. A comparison of traditional and Rasch cut points for assessing clinically important change in health-related quality of life among patients with asthma. *Qual. Life Res.* 15(10), 1639–1649 (2006).

▶36 de Vet HC, Terluin B, Knol DL *et al.* Three ways to quantify uncertainty in individually applied 'minimally important change' values. *J. Clin. Epidemiol.* 63(1), 37–45 (2010).

▶37 McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? *Qual. Life Res.* 4(4), 293–307 (1995).

▶38 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.* 52(9), 861–873 (1999).

▶39 Wyrwich K. Minimal important difference thresholds and the standard error of measurement: is there a connection? *J. Biopharm. Stat.* 14(1), 97–110 (2004).

▶40 Wyrwich KW, Tierney WM, Wolinsky FD. Using the standard error of measurement to identify important changes on the Asthma Quality of Life Questionnaire. *Qual. Life Res.* 11(1), 1–7 (2002).

▶41 Turner D, Schünemann HJ, Griffith LE *et al.* The minimal detectable change cannot reliably replace the minimal important difference. *J. Clin. Epidemiol.* 63(1), 28–36 (2010).

▶42 Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. *Med. Care* 27(3 Suppl.), S178–S189 (1989).

43 Cohen J. *Statistical Power Analysis for the Behavioural Sciences (2nd Edition)*. Lawrence Erlbaum Associates, NJ, USA (1988).

▶44 Farivar SS, Liu H, Hays RD. Half standard deviation estimate of the minimally important difference in HRQOL scores? *Expert Rev. Pharmacoeconomics Outcomes Res.* 4(5), 515–523 (2004).

▶45 Beaton DE. Simple as possible? Or too simple? Possible limits to the universality of the one half standard deviation. *Med. Care* 41(5), 593–596 (2003).

▶46 Wright JG. Interpreting health-related quality of life scores: the simple rule of seven may not be so simple. *Med. Care* 41(5), 597–598 (2003).

▶47 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.* 63(3), 270–281 (2010).

• **Provides empirical alternative to Cohen's arbitrary guidelines.**

▶48 Cocks K, King MT, Velikova G, Martyn St-James M, Fayers PM, Brown JM. 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 (EORTC QLQ-C30). *J. Clin. Oncol.* 29(1), 89–96 (2011).

• **Interesting presentation of expert opinion.**

49 Aaronson NK, Cull A, Kaasa S *et al.* The European Organisation for Research and Treatment of Cancer (EORTC) modular approach to quality of life assessment in oncology: an update. In: *Quality of Life and Pharmacoeconomics in Clinical Trials.* Spilker B (Ed.). Lippincott-Raven Publishers, PA, USA, 179–189 (1996).

▶50 Yost KJ, Eton DT. Combining distribution- and anchor-based approaches to determine minimally important differences: the FACIT experience. *Eval. Health Prof.* 28(2), 172–191 (2005).

▶51 Redelmeier DA, Guyatt GH, Goldstein RS. Assessing the minimal important difference in symptoms: a comparison of two techniques [see comments]. *J. Clin. Epidemiol.* 49(11), 1215–1219 (1996).

▶52 Deyo RA, Inui TS, Leininger J, Overman S. Physical and psychosocial function in rheumatoid arthritis. Clinical use of a self-administered health status instrument. *Arch. Intern. Med.* 142(5), 879–882 (1982).

▶53 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.* 84(4), 345–353 (2010).

▶54 Wyrwich KW, Metz SM, Kroenke K, Tierney WM, Babu AN, Wolinsky FD. Interpreting quality-of-life data: methods for community consensus in asthma. *Ann. Allergy Asthma Immunol.* 96(6), 826–833 (2006).

• **Exemplifies an innovative and complex methodology and a thorough analysis approach.**

55 Wyrwich KW, Metz SM, Kroenke K, Tierney WM, Babu AN, Wolinsky FD. Triangulating patient and clinician perspectives on clinically important differences in health-related quality of life among patients with heart disease. *Health Serv. Res.* 42(6 Pt 1), 2257–2274; discussion 2294–2323 (2007).

▶56 Wyrwich KW, Metz SM, Kroenke K, Tierney WM, Babu AN, Wolinsky FD. Measuring patient and clinician perspectives to evaluate change in health-related quality of life among patients with chronic obstructive pulmonary disease. *J. Gen. Intern. Med.* 22(2), 161–170 (2007).

▶57 King MT, Cella D, Osoba D *et al.* Meta-analysis provides evidence-based interpretation guidelines for the clinical significance of mean differences for the FACT-G, a cancer-specific quality of life questionnaire. *Patient Reported Outcome Measures* 2010(1), 119–126 (2010).

58 Fayers PM, Machin D. Sample sizes. In: *Quality Of Life: The Assessment, Analysis And Interpretation Of Patient-Reported Outcomes*. Wiley, Chichester, UK, 247–270 (2007).

59 Food and Drug Administration. Guidance for industry on patient-reported outcome measures: use in medical product development to support labeling claims. *Federal Register* 74(235), 65132–65133 (2009).

60 Dworkin RH, Turk DC, McDermott MP *et al.* Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. *J. Pain Symptom Manage.* 9(2), 105–121 (2008).

61 Ware JE, Keller SD. Interpreting general health measures. In: *Quality of Life and Pharmacoeconomics in Clinical Trials.* Spilker B (Ed.). Lippincott-Raven, NY, USA, 445–460 (1996).

▶62 Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR; Clinical Significance Consensus Meeting Group. Methods to explain the clinical significance of health status measures. *Mayo Clin. Proc.* 77(4), 371–383 (2002).

▶63 Beaton D, Boers M, Wells G. Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. *Curr. Opin. Rheumatol.* 14(2), 109–114 (2002).

•• **Thorough and thoughtful review.**

▶64 Hays R, Woolley J. The concept of clinically meaningful difference in health-related quality-of-life research. How meaningful is it? *Pharmacoeconomics* 18, 419–423 (2000).

•• **Insightful analysis of conceptual issues.**

65 Wyrwich KW, Spratt DI, Gass M, Yu H, Bobula JD. Identifying meaningful differences in vasomotor symptoms among menopausal women. *Menopause* 15(4 Pt 1), 698–705 (2008).

▶66 Sloan JA, Cella D, Frost M, Guyatt GH, Sprangers M, Symonds T; Clinical Significance Consensus Meeting Group. Assessing clinical significance in measuring oncology patient quality of life: introduction to the symposium, content overview, and definition of terms. *Mayo Clin. Proc.* 77(4), 367–370 (2002).

67 de Vet HC, Beckerman H, Terwee CB, Terluin B, Bouter LM. Definition of clinical differences. *J. Rheumatol.* 33(2), 434; author reply 435 (2006).

▶68 Beckerman H, Roebroeck ME, Lankhorst GJ, Becher JG, Bezemer PD, Verbeek AL. Smallest real difference, a link between reproducibility and responsiveness. *Qual. Life Res.* 10(7), 571–578 (2001).

▶69 Oken MM, Creech, RH, Tormey DC *et al.* Toxicity and response criteria of the Eastern Cooperative Oncology Group. *Am. J. Clin. Oncol.* 5, 649–655 (1982).

ELSEVIER

Journal of Clinical Epidemiology 61 (2008) 102−109

Journal of Clinical Epidemiology

REVIEW ARTICLES

Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes

Dennis Revicki[a,*], Ron D. Hays[b], David Cella[c], Jeff Sloan[d]

[a]*Center for Health Outcomes Research, United Biosource Corporation, 7101 Wisconsin Avenue, Suite 600, Bethesda, MD 20814, USA*
[b]*Division of General Internal Medicine and Health Services Research, University of California Los Angeles and RAND Corporation, Health Sciences Program, Los Angeles, CA, USA*
[c]*Center on Outcomes Research and Education, Evanston Northwestern Healthcare Research Institute and Northwestern University Feinberg School of Medicine, Evanston, IL, USA*
[d]*Section of Biostatistics, Mayo Clinic, Rochester, MN, USA*

Accepted 31 March 2007

Abstract

Objective: The objective of this review is to summarize recommendations on methods for evaluating responsiveness and minimal important difference (MID) for patient-reported outcome (PRO) measures.

Study Design and Setting: We review, summarize, and integrate information on issues and methods for evaluating responsiveness and determining MID estimates for PRO measures. Recommendations are made on best-practice methods for evaluating responsiveness and MID.

Results: The MID for a PRO instrument is not an immutable characteristic, but may vary by population and context, and no one MID may be valid for all study applications. MID estimates should be based on multiple approaches and triangulation of methods. Anchor-based methods applying various relevant patient-rated, clinician-rated, and disease-specific variables provide primary and meaningful estimates of an instrument's MID. Results for the PRO measures from clinical trials can also provide insight into observed effects based on treatment comparisons and should be used to help determine MID. Distribution-based methods can support estimates from anchor-based approaches and can be used in situations where anchor-based estimates are unavailable.

Conclusion: We recommend that the MID is based primarily on relevant patient-based and clinical anchors, with clinical trial experience used to further inform understanding of MID.

Keywords: Patient-reported outcomes; Health-related quality of life; Minimal important differences; Clinical significance; Anchor-based methods; Distribution-based methods

1. Introduction

Patient-reported outcomes (PROs) are frequently incorporated in clinical trials comparing health interventions for chronic diseases. These PROs include measures of health-related quality of life (HRQL), symptoms, and treatment satisfaction. PROs provide the patient's perspective and help us understand the effects of disease and treatment on symptoms, functioning, and other outcomes [1−3]. For many chronic diseases, PROs represent one of the most important health outcomes for evaluating the effectiveness of treatments and changes in disease trajectory. As far back as Hippocrates, listening to the patient has been considered an integral part of medical science [4]. Therefore, the patient's perspective of her health is integral to understanding health outcomes. The application of relevant and psychometrically sound PROs in clinical trials assists patients, their family members, and clinicians in understanding the comprehensive impact of treatment on patient symptoms, functioning, treatment preferences, and general well being.

To be useful in clinical trials evaluating new health interventions, PROs, similar to other health outcomes, must have acceptable reliability and validity [1,2,5,6]. Responsiveness is an aspect of construct validity [7] and is determined by evaluating the relationship between changes in clinical and patient-based endpoints and changes in the PRO scores over time, or based on the application of a treatment of known and demonstrated efficacy [2,5,8]. Responsiveness can be evaluated based on observational studies or in clinical trials. Evidence supporting responsiveness and

* Corresponding author. Tel.: +301-654-9729; fax: +301-654-9864.
E-mail address: dennis.revicki@unitedbiosource.com (D. Revicki).

0895-4356/08/$ − see front matter
doi: 10.1016/j.jclinepi.2007.03.012

What is new?

- Recommend that the minimal important difference (MID) be based primarily on appropriate patient-based and clinical anchors that are correlated at ⩾0.30 with the patient-reported outcome (PRO), with clinical trial experience used to further inform understanding of MID.
- MID may vary by population and context, and thus a single MID may be insufficient for all study applications involving a PRO instrument.
- Estimation of MID for a specific PRO measure should be based on multiple approaches and triangulation of methods.
- Various methods for estimating MIDs often converge, and generalizability of MID estimates for similar applications is supported.
- Recommend basing the final selection of MID values on systematic review and evaluation process such as a modified Delphi method.

for interpreting PRO results is critical for clinical trial settings. Information on the interpretation of changes or differences in PRO scores is based on the minimal important difference (MID). Demonstrating a MID is also important evidence for achieving successful PRO claims through regulatory agencies [9,10]. Nonetheless, virtually all instruments found to differentiate among clinically distinct groups are also found to be responsive to change.

Although responsiveness and interpretation of PRO measures have been discussed for the past 15 years or more [2,5,7,11−29], recommendations about the best approach for evaluating responsiveness and determining MIDs for PRO instruments are still needed. For example, the FDA requested further information and guidance on methods for determining responsiveness and MID [9]. Although there is an evolving consensus as to the best approach to evaluating responsiveness and MID [25,26,29], there is no clear statement about the recommended methods and about important issues underlying responsiveness and MID.

This report focuses on issues and recommendations for evaluating responsiveness and MID for PRO measures in chronic disease. These issues are especially germane given that for most chronic diseases cure is not feasible, and that the main objective of treatment is to maintain or improve patient functioning and well being. The remainder of this report will cover (1) conceptual issues and definitions; (2) methods for evaluating responsiveness and MID; (3) recommended decision criteria for determining MID; and (4) summary and conclusions. We will illustrate methods and concepts using published health outcomes literature.

2. Interpretation of PROs: conceptual issues and definitions

PROs require the patient to assign a response to questions (or statements) about their perceptions or activities, such as symptoms, capabilities, or performance of roles or responsibilities. These responses are typically combined in some way to create summary scores that can be used to measure concepts such as physical, psychological, or social functioning and well being, or symptom burden or severity. Symptoms can be rated based on frequency, severity, duration, degree of bother, or impact on patient activities. Demonstrating the ability to detect responsiveness to meaningful change is necessary but not sufficient for estimating the smallest change in score that can be regarded as important. This amount of change score has been referred to as the MID, and when connected to clinical anchors, sometimes as the minimal clinically important difference (MCID). Responsiveness represents the instrument's ability to detect changes whereas the MID denotes the smallest score or change in score that would likely be important from the patient's or clinician's perspective.

Because responsiveness and MID depend on population and contextual characteristics, there is not necessarily a single MID value for a PRO instrument across all applications and patient samples. There is often a range in MID estimates that varies across patient population and clinical study context.

The MID has been defined as the smallest difference in scores of a PRO measure that is perceived by patients as beneficial or harmful, and which would lead the clinician to consider a change in treatment [8,22]. A number of anchor-based and distribution-based methods have been used to determine the MID for PRO measures [22,23,25]. However, the current situation for determining the MID is fluid and evolving, and there is no clear consensus as to the recommended, best-practice approach for determining the MID [22]. Some have recommended to estimate the MID based on several anchor-based methods, with relevant clinical or patient-based indicators, and to examine various distribution-based estimates (i.e., effect size, standardized response mean, standard error of measurement [SEM]) as supportive information, and then to triangulate on a single value or small range of values for the MID [28−31]. Similar to virtually every measure used in medicine, all PRO assessments include some measurement error. The amount of error in most well-developed PRO instruments is similar to or less than what is observed in most standard clinical measures [32]. One needs confidence that observed changes in scores over time with treatment are not primarily attributable to error. Put another way, instrument reliability helps ensure that observed differences between treatment groups exceed what one might expect based upon measurement error alone [21,32,33]. Confidence in a specific MID value evolves over time and is confirmed by additional research evidence.

2.1. Minimal versus meaningful differences

The idea of a MID came about in the literature somewhat indirectly. In efforts to evaluate an asthma quality of life questionnaire, Jaeschke et al. [11,34] attempted to estimate how much of a difference in scores would result in some change in clinical management that is to be considered clinically meaningful. They ''developed an approach to elucidating the significance of changes in score in quality of life instruments by comparing them to global ratings of change. Using this approach (they) established a plausible range within which the minimal clinically important difference (MCID) falls'' [11]. It is important to note that Jaeschke did not define an MCID as much as indicating that the technique they used was a reasonable approach to produce an estimate of the interval within which an MCID falls. They concluded that such an approach provided support for an estimate for a clinically meaningful difference that could be applied to both groups and individual patients. They, however, did not claim to actually derive a truly ''minimal'' difference estimate.

More recently, Sloan proposed a distribution-based method (the empirical rule effect size or ERES method) from a perspective of finding an effect size that was nonignorable [35,36]. Sloan noted that many of the competing methods seemed to converge into the same general area in terms of proportion of standard deviation (SD) of the PRO instrument under study [37,38]. The MCID derived by Jaeschke et al. [11,34], for example, was precisely equal to the ½SD estimate produced via the ERES method. Further data by Norman et al. [24] seemed to confirm this hypothesis that ½SD could indeed be a minimally important difference in some circumstances. Sloan was not searching for a *minimum* but was instead looking for a conservative estimate for a clinically meaningful difference (i.e., an ''obviously'' important difference). Recent work has indicated that there are situations where a difference smaller than ½SD may be meaningful [27,36]. Evidence supporting any MID is needed to justify such estimates, and sensitivity analysis using multiple approaches is recommended wherever possible. Where no such evidence is available, the ½SD estimate may be a reasonable place to start as a meaningful difference.

The idea of finding an estimate of the ''minimum'' important difference is at once intuitively appealing and mathematically challenging. Some have proposed alternative terminology (summarized in Sloan et al. [36]), but the nomenclature is still inconsistent. Can we ever get to a true minimum? It is more reasonable to assume that we can get to a consensus on what is meaningful for practical purposes and use this as a benchmark. Determining when a minimum has been achieved is harder to gauge and requires greater precision. Thus, the MID may only be useful to sort cases as improved or not (or as worsened or not), with the substantial benefit judged based on a sizable increase in the proportion of improved cases (or sizable decrease in the number of declined cases) observed by treatment group.

3. Methods of evaluating responsiveness and clinical significance

Longitudinal studies are needed to determine whether a PRO instrument is responsive to changes or differences. These studies may be randomized clinical trials comparing treatments of known efficacy or observational studies where patients are treated with usual medical care and followed over relevant periods of time. For clinical trial designs, there needs to be some evidence that the treatment is effective and that the expected changes in clinical status are linked to expected changes in the PRO measure. Careful attention to selecting treatments for clinical trials will attenuate circumstances where the researcher may falsely conclude that the PRO is not response, because the PRO and clinical endpoints are not associated with each other. To assess responsiveness, some criterion is needed to identify whether patients have changed (either improved or worsened) over time. These criteria, or anchors, may be clinical endpoints, patient-rated global improvement, change in other PRO measures, or some combination of clinical and patient-based outcomes. In addition to the anchor-based approaches, responsiveness and clinical significance can be informed by previous work using distribution-based methods and through systematic reviews of clinical trials.

3.1. Anchor-based methods

The anchor-based approaches use an external indicator, either clinical or patient-based, to assign subjects into several groupings reflecting no change, small positive changes, large positive changes, small negative changes, or large negative changes in clinical or health status. The anchors can be clinical (i.e., laboratory measures, physiological measures, and clinician ratings) or patient based, such as global ratings of change or actual changes in PRO measures that have demonstrated MID in the target patient population. It is strongly recommended to use multiple independent anchors and to examine and confirm responsiveness across multiple samples [22,25,28,29,39,40].

Selecting anchors should be based on criteria of relevance for the disease indication, clinical acceptance, and validity and evidence that the anchors have some relationship with the PRO measure. The best anchors for estimating the MID—whether retrospective measure of change, knowledge about the course of health over time, or clinical parameters—are ones that identify those who have changed to a small but meaningful degree. Including individuals who change beyond a small but meaningful degree risks over-estimating the MID. It is important to identify the subset of people who have experienced minimal change. Those patients who have changed by a minimal amount have been

identified by asking study participants at follow-up to report how much they changed since baseline of a study using a multiple categorical response scale. People who reported either getting *a little better* or *a little worse* constitute the minimal change subgroup. The change in PRO measures reported by this subgroup is the estimate of the MID as perceived by the patient. One can decide to examine change for those getting worse versus getting better separately or pool them together after accounting for the difference in the direction of change.

Retrospective self-reports (such as health transition questions) are known to be subject to recall bias [41]. When retrospective change items are used as anchors, it is useful to determine if they reflect the baseline (pretest) and present (posttest) status equally. In theory, retrospective change items should correlate positively with the posttest and have a negative correlation of equal magnitude with the pretest as illustrated in the following formulas: $r(x, y - x) = r(x,y)$ and $r(y, y - x) = r(y,-x) = -r(x,y)$, where $r(.,.)$ is the correlation, x is the pretest, and y is the posttest. In reality, retrospective self-reports tend to correlate more strongly with the posttest than they do with the pretest because current status unduly influences the retrospective perception of change. For example, Walters and Brazier [42] found moderate correlations (mean 0.45, range: 0.18−0.57) between responses to a retrospective measure of global change and the SF-6D at follow-up across nine studies. Correlations with initial assessments were systematically lower (mean 0.22, range: 0.01−0.41). Thus, these correlations should be interpreted with flexibility and allowance for lack of equality.

For a clinical parameter, it is also necessary to establish the amount of change on the anchor that is a reasonable indicator of minimal change. Estimating the MID requires agreement about what constitutes a minimal change in the anchor. Kosinski et al. [39] defined minimal improvement on their clinical measures as 1%−20% improvement in the number of swollen and tender joints in a study of patients with rheumatoid arthritis. Although this may be a reasonable threshold, other investigators might argue for another threshold (e.g., >10%; >20% improvement). Any anchor that is chosen should have a ''nontrivial'' association with change in the PRO measure. If the correlation between the anchor and PRO change is zero, then the anchor is not useful for establishing the MID. While a nontrivial correlation is important, a clinical anchor cannot ''hope to capture the richness and variation of the construct of HRQL'' [17]. Using Cohen's [43] rules of thumb, we recommend 0.30−0.35 as a correlation threshold to define an acceptable association between an anchor and a PRO change score, although alternative thresholds may be acceptable in the presence of supplementary information.

The variety of possible anchors and uncertainty in the anchor cut-point that defines a minimal difference make a single estimate of MID problematic. Using the retrospective report anchor as an example, the recall item might refer globally to change in ''health,'' ''health-related quality of life,'' or ''quality of life.'' Moreover, the anchor might be worded more specifically such as ''physical functioning,'' ''pain,'' etc. The choice of words can lead to variability in the performance of the anchor. Any specific anchor may be more or less appropriate for different PRO domains. For example, an energy/fatigue scale might be expected to change more than a pain scale in response to change in hematocrit [44].

There also needs to be an understanding of the trajectory of health outcomes in the target disease to evaluate responsiveness. For example, do most patients improve over time with treatment, as with seasonal allergic rhinitis, or as in many chronic diseases (e.g., COPD, arthritis, etc.), is the expected trajectory one of maintenance or varying levels of deterioration in health status over time, even with treatment? Other factors that can lead to variation in the estimation of the MID include whether the people being evaluated are high or low on the measure at baseline, whether they improve or decline in HRQL over time, and whether they have similar demographic, clinical, and other characteristics [18].

Once groups of patients are identified as improving, worsening, or remaining stable based on several relevant external anchors, several data analyses and indicators can be used to examine responsiveness. First, analysis of variance or covariance procedures can be performed comparing differences in mean baseline to endpoint changes in the PRO scores across the meaningful change groups (i.e., stable versus small improvement, stable versus moderate improvement, etc.). Second, responsiveness to change is frequently evaluated using different indicators [5,33], such as the effect size (ES) [45], standard response mean (SRM) [46], and responsiveness statistic (RS) [8].

3.2. Previous clinical trial experience

As the medical literature on PROs applied in clinical trials increases [47], it is increasingly possible to understand responsiveness and MID for different PRO instruments based on demonstrated differences between active and placebo treatments or between two or more active treatments in clinical trials. Systematic reviews of the clinical trial literature can therefore be used to determine clinical significance. For example, Niebauer et al. [48] in a systematic review found consistent evidence that omalizumab resulted in a 0.30 point improvement (i.e., 0.30 effect size) in asthma quality of life questionnaire scores when used as add-on therapy in moderate to severe asthma. There are observations that the MID determined through anchor-based methods seen in observational, psychometric evaluation studies may differ from MIDs seen in randomized clinical trials (D. Patrick, personal communication, February 23, 2006). As evidence accumulates in clinical trials, the observed changes or difference in PRO measures based on effective treatments provides a rich and valuable source of data on responsiveness and clinical significance. For example, Jones [49] summarized the clinical trial experience for

the St. Georges Respiratory Questionnaire (SGRQ) in the chronic obstructive pulmonary disease literature. He found that across eight different studies, the recommended MID for the SGRQ total score was approximately 4 points (i.e., 0.32 effect size).

Therefore, it is recommended that the previous clinical trial literature for the targeted indication be reviewed and synthesized to identify evidence on responsiveness and MID for the selected PRO instruments. These PRO data provide an excellent resource for understanding the application and performance of PRO measures in the clinical trial environment and can be used to further support the evidence base on responsiveness and MID for interpreting PRO data.

3.3. Distribution-based methods

Distribution-based methods convey a notion that a MID can be estimated based on the distribution of observed scores in a relevant sample. Guyatt et al. [22] provide concise and complete exposition of the various distributional methods. Some have criticized distribution-based methods because they are anchor free (i.e., ''meaning-free''). Still others have expressed concern over accepting a ''purely'' statistical argument to support the choice of a MID, which is interesting when statistical theory underlies virtually all clinical investigation [50].

Some researchers have suggested that the ½SD estimate [24] or that the SEM [51,52] may approximate a MID for some PRO instruments. Although this magnitude of change is certainly clinically significant and meaningful, it is not necessarily minimal. Empirical evidence from previous studies, physiological arguments, and statistical theory shows a tendency to converge to the ½SD criteria as being meaningful to patients [40]. Although different distribution-based indicators demonstrate that change has occurred and provide some insight as to whether the change (responsiveness) is small or large, the indices do not necessarily inform as to whether the observed change is MID. To determine MID, it is necessary to get information as to whether the observed change is important from the patient's or clinician's perspective [40]. MIDs have been observed, however, to be as small as 0.25 to 0.33 ES (or SD units) [27,29].

The distribution-based indices provide no *direct* information about the MID. They are simply a way of expressing the observed change in a standardized metric. This makes it possible to compare change observed for measures that have a different raw metric and the degree of deviation (individual and group level) within the sample. ES estimates can be compared to Cohen's guidelines about the magnitude, but anchor-based methods are the only way to estimate the MID directly. The SEM has been proposed as a method of relevance to MID estimation. This suggestion is based on anecdotal observations that the SEM was approximately equal to the estimated MID [51]. Norman et al. [24] note that 1 SEM is approximately the same as a 0.5 difference on a seven-point scale and 1 SEM is approximately ½SD when the reliability is 0.75. But why should 1 SEM have anything to do with the MID? The SEM measurement is estimated by the product of the SD and the square root of 1-reliability of a measure. The SEM is used to set the confidence interval (CI) around an individual score, that is, the observed score plus or minus 1.96 SEMs constitutes the 95% CI. In fact, the reliable change index proposed earlier by Jacobson and Truax [12] is based on defining change using the statistical convention of exceeding 2 standard errors.

Distribution-based methods are most applicable when the goal of estimating a clinically meaningful difference does not have a heavy reliance on the estimate needing to be minimal. There is a concern about the use of a conservative estimate for MID as it may require setting a criteria for success beyond what is achievable for a given treatment. Therefore, we recommend that the distribution-based measures are used as supportive information for MID estimates from different anchor-based approaches and systematic reviews of the clinical trial literature.

3.4. Determining the MID for PRO instruments

For interpreting differences or changes in PRO instruments, information needs to be provided as to whether the changes seen in the scores are important from either the patient's or clinician's perspective. The clinical meaningfulness of the observed change is based on that change perceived as minimally important and as beneficial or harmful from the patient's viewpoint. It is recommended that the patient's perspective be given the most weight, because these are PROs, although the clinician's perspective is considered important as well. The MID is best estimated using multiple anchors with the same external criteria used to evaluate responsiveness of the PRO measure. However, there are differences in how these data are used and compared to determine MID. Because the focus is on determining the MID, it is necessary to identify the smallest difference or change that is important to the patient.

In many cases, global assessments of change in health or clinical status are used to categorize patients into groups that reflect, based on their own reports, different amounts of change in the construct of interest. Most often the MID is determined as the change observed in the small improvement group, as long as the small improvement group demonstrates changes that are larger than the stable group. If the changes are similar, there is uncertainty about the group definition, change in the PRO measures, and the MID. There are cases where there is some variation observed even among the stable group, in these cases it is informative to examine the difference in mean baseline to endpoint change scores between the stable group and the small improvement (or worsening) group.

Note that there is evidence that there is asymmetry in worsening and improvement in PROs depending on the

specific disease [20,27]. Clinician global assessments of change in clinical status or evaluations of clinical severity, clinical response criteria (i.e., American College of Rheumatology [ACR] response criteria), or other indicators can be used to determine MID. For these clinical anchors, it will be necessary to identify, based on previous research or clinical consensus, what a small and clinically meaningful effect may be based on these measures. For example, in rheumatoid arthritis, the differences between groups of stable patients and those experiencing a 20% ACR response can be used to determine the MID of a PRO score. If multiple anchors are used, there will be several different estimates of MID derived corresponding to these different anchors, and the result will be a range of MID estimates for the targeted PRO instrument.

4. Recommended decision criteria for determining MID

The application of multiple methods to determine the MID for a PRO instrument in a specific patient population will almost always result in a *range* of values for the MID. This is the essence of triangulation, that is, examining multiple values from different approaches and hopefully converging on a small range of values (or one single value). It is recommended that the different MID estimates be graphed to visually depict the range of estimates. Figure 1 provides a summary of MID estimates from a study by Yost et al. [27]. To identify a single MID value (or narrow range of MID values), it is recommended that the anchor-based estimates be assigned the most weight, and experience from clinical trials be used to further support and perhaps further narrow the range of values. Interpretation of the MID from different anchors should also take into account the proximity of the anchor to the target PRO measure, that is, assign more importance to MIDs generated from more closely linked concepts.

A systematic consensus process involving several clinicians and health outcome researchers is recommended and can be completed, based on Delphi methods, to arrive at a single MID value, or at least a narrower range of values. There is no consensus as to how much data are needed as supportive evidence for the MID of a PRO instrument. Clearly, the more data and evidence the better, but a single, generalizable study with multiple patient-based and clinical anchors may be sufficient. As with other aspects of construct validity, responsiveness and the MID value are confirmed based on accumulating evidence from multiple studies and, with additional data, we can be more confident in the MID value. It would be rather unusual for a single MID to be appropriate for all applications and across all patient populations. For example, the MID derived for an asthma-specific quality of life measure in mild to moderate asthma patients may not be generalizable to clinical trials comparing an add-on treatment for patients with moderate to severe asthma [48].

5. Summary and conclusions

For PRO endpoint data to be accepted as evidence of treatment efficacy there must be evidence documenting the instrument's conceptual framework, content validity, and psychometric qualities. For responsiveness, it is necessary to demonstrate that the PRO scores are sensitive to actual changes in health status. Although demonstrating

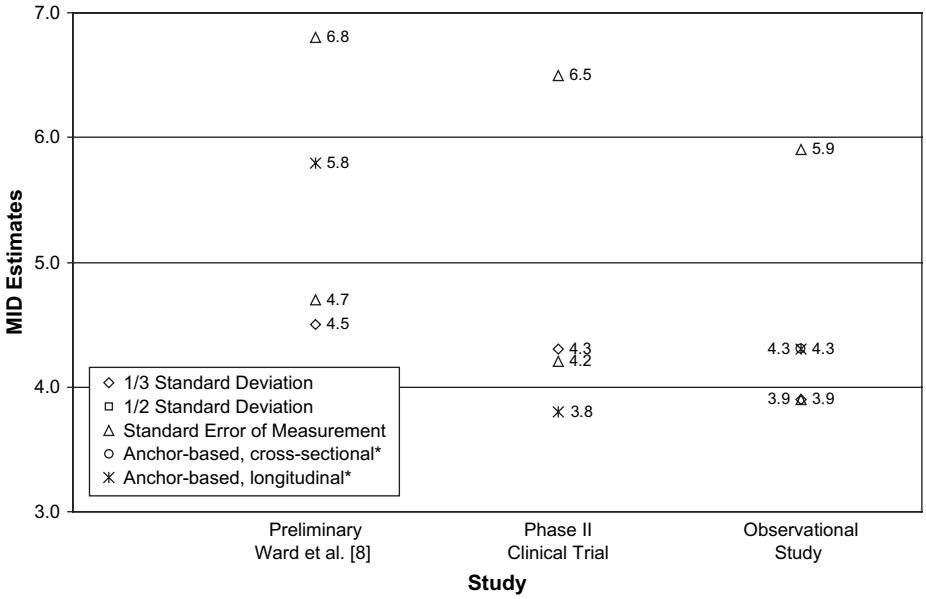


Fig. 1. Summary of Distribution- and Anchor-Based Estimates of Minimally Important Differences (MIDs) for the Trial Outcome Index-Colorectal (TOI-C). Only anchor-based values with effect sizes between 0.2 and 0.5 were plotted. *If more than one value was available for a given type of estimate, the average was plotted.

responsiveness is a key component to establishing an instrument's construct validity, it is also important to determine the MID to assist in interpreting statistically significant PRO results in clinical trials. In addition, the MID for a PRO instrument that is specified as a primary or important secondary endpoint is clearly useful for calculating statistical power and for determining sample sizes for clinical trials. The MID may vary by population and context, and no one MID may be valid for all study applications involving a PRO instrument. Responsiveness and MID must be demonstrated and documented for the particular study population.

The estimation of MID for a specific PRO measure should be based on multiple approaches [22,40] and triangulation of methods. Anchor-based methods applying various relevant patient-rated, clinician-rated, and disease-specific variables provide primary and meaningful estimates of an instrument's MID. Previous results, including the PRO measures, from clinical trials can also provide insight into observed effects based on treatment comparisons, and these data should be used to help determine MID. Distribution-based methods can support and help interpret estimates from anchor-based approaches and can be used in situations where anchor-based estimates are unavailable. We recommend that the MID be based primarily on relevant patient-based and clinical anchors, with clinical trial experience used to further inform understanding of MID.

Multiple approaches to estimating the MID will produce a range of different values, and decision guidance is needed to select a single value or narrow range of MID values. Based on examination of the resultant MID estimates, it is often possible to select a narrow range of MID estimates. When this is difficult or there is some uncertainty about which MID may be best, we recommend basing the final selection of MID values on some systematic review and evaluation process such as a modified Delphi method.

The MID for a PRO instrument is not an immutable characteristic; it may vary across populations and treatments. Incremental changes in the PRO measure may be influenced by disease severity and the treatment context. Continuing experience with the PRO instrument in different population contexts and clinical trial situations will inform the understanding of MID for the instrument. As with construct validity, accumulating evidence across multiple studies will help clinicians, and health outcomes researchers gain confidence in interpreting important differences for the PRO measure.

Selecting a MID estimate for clinical trial planning or for interpretation of PRO endpoint findings should be based on the existing knowledge on MID for the specific PRO instrument. There needs to be a clear rationale for the selection of the MID value especially in situations where there may be range of MID estimates for the instrument. It is best to select the value based on clinically relevant anchors that are proximal to the concept(s) measured by the PRO instrument, and based on the understanding of the disease area and patient population. For interpreting PRO results from a clinical trial, the proposed MID value should be identified a priori in the PRO statistical analysis plan. The rationale for selecting the MID value for a study should be clear and straightforward based on the existing evidence. If there is uncertainty about this value, additional supportive research may be needed to strengthen the evidence base.

PRO measures provide the patient's perspective on the impact of disease and treatment. These health outcomes allow for a more comprehensive evaluation of a medical intervention and require evidence on sound psychometric characteristics, including responsiveness and interpretation guidelines. MID estimates, if based on systematic research and relevant anchors, provide the basis for interpreting clinical trial results and help regulatory agencies, clinicians, and patients understand the effects of treatment of symptoms, and patient functioning and well being.

Acknowledgments

This paper was supported in part by Genentech, South San Francisco, California, the UCLA/DREW Project EXPORT, National Institutes of Health, National Center on Minority Health & Health Disparities, (P20-MD00148-01), the UCLA Center for Health Improvement in Minority Elders/Resource, Centers for Minority Aging Research, National Institutes of Health, National Institute of Aging, (AG-02-004), and the National Institute of Aging (AG20679-01).

References

[1] Leidy NK, Revicki DA, Geneste B. Recommendations for evaluating the validity of quality of life claims for labeling and promotion. Value Health 1999;2:113−27.

[2] Revicki DA, Osoba D, Fairclough D, Barofsky I, Berzon R, Leidy NK, et al. Recommendations on health-related quality of life research to support labeling and promotional claims in the United States. Qual Life Res 2000;9:887−900.

[3] Wilke RJ, Burke LB, Erickson P. Measuring treatment impact: a review of patient-reported outcomes and other efficacy endpoints in approved product labels. Contr Clin Trials 2004;25:535−52.

[4] Hippocrates. On decorum. Hippocrates, with an English translation. In: Jones WH, translator. Cambridge, MA: Harvard University Press; 1923;267−301.

[5] Hays RD, Revicki DA. Reliability and validity (including responsiveness). In: Fayers P, Hays R, editors. Assessing quality of life in clinical trials. 2nd edition. New York: Oxford University Press; 2005.

[6] Fayers PM, Machin D. Quality of life: Assessment, analysis and interpretation. Chichester: John Wiley & Sons; 2000.

[7] Hays RD, Hadorn D. Responsiveness to change: an aspect of validity, not a separate dimension. Qual Life Res 1992;1:73−5.

[8] Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chronic Dis 1987;40:171−8.

[9] FDA. Guidance for industry—patient-reported outcome measures: Use in medical product development to support labeling claims. Silver Spring, MD: FDA; 2006.

[10] Committee for Medicinal Products for Human use. Reflection paper on the regulatory guidance for the use of health-related quality of life

(HRQL) measures in the evaluation of medicinal products. London: EMEA; 2005.
[11] Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Contr Clin Trials 1989;10:407−15.
[12] Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. J Clin Consult Psychol 1991;59:12−9.
[13] Liang MJ. Evaluating measurement responsiveness. J Rheumatol 1995;22:1191−2.
[14] Lydick E, Epstein RS. Interpretation of quality of life changes. Qual Life Res 1993;2:221−6.
[15] Testa M, Lenderking WR. Interpreting pharmacoeconomic and quality-of-life clinical trial data for use in therapeutics. Pharmacoeconomics 1992;2:107−17.
[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: 139−44.
[17] Lydick E, Yawn BP. Clinical interpretation of health-related quality of life data. In: Staquet M, Hays R, Fayers P, editors. Quality of life assessment in clinical trials: Methods and practice. Oxford: Oxford University Press; 1998.
[18] Hays RD, Woolley JM. The concept of clinically meaningful difference in health-related quality-of-life research. How meaningful is it? Pharmacoeconomics 2000;18:419−23.
[19] Beaton DE, Bombardier C, Katz JN, et al. Looking for important changes/differences in studies of responsiveness. J Rheumatol 2001;28:400−5.
[20] Cella D, Hahn EA, Dineen K. Meaningful changes in cancer-specific quality of life scores: differences between improvement and worsening. Qual Life Res 2002;11:207−21.
[21] Cella D, Bullinger M, Scott C, Barofsky I, Sloan JA. Group vs individual approaches to understanding the clinical significance of differences or changes in quality of life. Mayo Clin Proc 2002;77:384−92.
[22] Guyatt G, 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.
[23] Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol 2003;56: 395−407.
[24] 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;41:582−92.
[25] Wyrwich KW, Bullinger M, Aaronson N, Hays RD, Patrick DL, Symonds T, et al. Estimating clinically significant differences in quality of life outcomes. Qual Life Res 2005;14:285−95.
[26] Sloan JA, Cella D, Hays RD. Clinical significance of patient-reported questionnaire data: another step toward consensus. J Clin Epidemiol 2005;58:1217−9.
[27] Yost KJ, Cella D, Chawla A, Holmgren E, Eton DT, Ayanian JZ, et al. Minimally important differences were estimated for the Functional Assessment of Cancer Therapy-Colorectal (FACT-C) instrument using a combination of distribution- and anchor-based approaches. J Clin Epidemiol 2005;58:1241−51.
[28] Yost KJ, Eton DT. Combining distribution- and anchor-based approaches to determine minimally important differences: the FACIT experience. Eval Health Prof 2005;28:172−91.
[29] Revicki DA, Erickson P, Sloan J, Dueck A, Guess H, Santanello N. Interpreting and reporting results based on patient-reported outcomes. Value in Health (in press).
[30] Cella D, Eton DT, Lai JS, Peterman AH, Merkel DE. Combining anchor and distribution based methods to derive minimal clinically important differences on the Functional Assessment of Cancer Therapy anemia and fatigue scales. J Pain Symptom Manage 2002;24:547−61.
[31] Hays RD, Brodsky M, Johnston MF, Spritzer KL, Hui K. Evaluating the statistical significance of health-related quality of life change in individual patients. Eval Health Prof 2005;28:160−71.
[32] Hahn EA, Chassany O, Fairclough D, Hays RD, Wong G, Cella D. A guide for clinicians to compare the accuracy and precision of health-related quality of life data relative to other clinical measures. Mayo Clin Proc [in press].
[33] Sprangers MAG, Moinpour CM, Moyniyhan TJ, Patrick DL, Revicki DA. Assessing meaningful changes in quality of life over time: a user's guide for clinicians. Mayo Clin Proc 2002;77:561−71.
[34] Juniper EF, Guyatt GH, Willan A, Griffith LE. Determining a minimal important chance in the disease-specific quality of life questionnaire. J Clin Epidemiol 1994;47:81−7.
[35] Sloan JA, Vargas-Chanes D, Kamath CC, Sargent DJ, Novotny PJ, Atherton P, et al. Detecting worms, ducks and elephants: a simple approach for defining clinically relevant effects in quality-of-life measures. J Cancer Integr Med 2003;1:41−7.
[36] Sloan JA, Frost MH, Halyard MH, Dueck A, Atherton P, Novotny P, et al. Applying QOL assessments: solutions for oncology clinical practice and research, part 1. Curr Probl Cancer 2005;29:271−351.
[37] Sloan J, Symonds T, Vargas-Chanes D, Fridley B. Practical guidelines for assessing the clinical significance of health-related quality of life changes within clinical trials. Drug Inf J 2003;37:23−31.
[38] Sloan JA. Assessing the minimally clinically significant difference: scientific considerations, challenges and solutions. J Chronic Obstructive Pulmonary Dis 2005;2:57−62.
[39] Kosinski M, Zhao SZ, Dedhiya S, Osterhaus JT, Ware JE. Determining the minimally important changes in generic and disease-specific health-related quality of life questionnaires in clinical trials of rheumatoid arthritis. Arthritis Rheum 2000;43:1478−87.
[40] Osoba D. The clinical value and meaning of health-related quality-of-life outcomes in oncology. In: Lipscomb J, Gotay CC, Snyder C, editors. Outcomes assessment in cancer: Measures, methods, and applications. Cambridge: Cambridge University Press; 2005:386−405.
[41] Schwartz N, Sudman S. Autobiographical memory and the validity of retrospective reports. New York: Springer-Verlag; 1994.
[42] Walters SJ, Brazier JE. What is the relationship between the minimally important difference and health state utility values? The case of the SF-6D. Health QOL Outcomes 2003;1:4.
[43] Cohen J. Statistical power analysis for the behavioral sciences. 2nd edition. Hillsdale, NJ: Lawrence Earlbaum Associates; 1988.
[44] Beusterien KM, Nissenson AR, Port FK, Kelly M, Steinwald B, Ware JE. The effects of recombinant human erythropoietin on functional health and well-being in chronic dialysis patients. J Am Soc Nephrol 1996;7:763−73.
[45] Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Med Care 1989;27:S178−89.
[46] Liang MJ, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. Med Care 1990;28:632−42.
[47] Kaplan RM. Measuring quality of life for policy analysis: past, present, and future. In: Lenderking WR, Revicki DA, editors. Advancing health outcomes research methods and clinical applications. McLean, VA: International Society for Quality of Life Research; 2005:1−35.
[48] Niebauer K, Dewilde S, Fox-Rushby J, Revicki DA. Impact of omalizumab on quality-of-life outcomes in patients with moderate-to-severe allergic asthma. Ann Allergy Asthma Immunol 2006;96: 316−26.
[49] Jones PW. Interpreting thresholds for a clinically significant changes in health status in asthma and COPD. Eur Respir J 2002;19: 398−404.
[50] Sloan JA, Dueck A. Issues for statisticians in conducting analyses and translating results for quality of life end points in clinical trials. J Biopharm Stat 2004;14:73−96.
[51] Wyrwich KW, Nienaber N, Tierney W, Wolinsky F. Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life. Med Care 1999;37:469−78.
[52] Wyrwich KW, Tierney W, Wolinsky F. Further evidence supporting an SEM-based criteria for identifying meaningful intra-individual changes in health-related quality of life. J Clin Epidemiol 1999;52: 861−73.

VALUE IN HEALTH 15 (2012) 401–403

Available online at www.sciencedirect.com

SciVerse ScienceDirect

journal homepage: www.elsevier.com/locate/jval

EDITORIAL

Beyond the FDA PRO Guidance: Steps toward Integrating Meaningful Patient-Reported Outcomes into Regulatory Trials and US Drug Labels

When sitting with a patient—or as a patient—deciding whether to start a treatment, often the first question asked is, "how will it make me feel" or "how have others like me felt." Yet this information is conspicuously absent from most US drug labels and published results of regulatory clinical trials.

The guiding principle here is that the patient perspective, which is usually best captured via a patient-reported outcome (PRO) measure, is *always* relevant and should be assessed in all pivotal clinical trials unless the impact of a product on the patient experience is already well known. Even if a product is expected to have little or no impact on how a patient feels, substantiating that expectation with data is informative to decision makers. Patient-reported information may reflect symptomatic benefits or symptomatic toxicities of a product or may demonstrate impact on the overall patient experience measured as health-related quality of life (HRQOL). Arguably, not including such information in a trial or label represents an omission that results in decision makers having incomplete information to balance risks with benefits.

Why is this information so often missing from labels? Is it that sponsors simply do not measure patients' symptoms or HRQOL? Or perhaps sponsors do collect this information but the Food and Drug Administration (FDA) feels their approaches are methodologically inadequate to merit inclusion in labels? Or maybe it is overly challenging or infeasible to collect and analyze patient reports compared with survival-based or surrogate end points.

An informative new article in the current issue of *Value in Health* [1] provides descriptive data suggesting that all the above reasons may contribute, but, as described below, are surmountable.

Current status: The sponsor

While industry sponsors do include PRO measures in many studies, these are often generic tools used to enable economic analyses by European regulators or to explore HRQOL and nonspecific symptoms. Outcomes and measures are frequently selected late in a development cycle when it is too late to conduct qualitative work to establish which outcomes are most important in the target population, or to assess how measures perform. Statistical power is rarely reserved for PRO analyses in pivotal trials. A disconnect between clinical investigators and PRO experts lies within most companies; this limits the possibility that PROs will be prominent in a study design. PRO experts tend to be associated more closely with postmarketing research units than with preapproval clinical development teams, and therefore PROs are more frequently integrated into observational research following approval to help guide marketing strategy. But this information often is not published or is unavailable to regulators, patients, clinicians, or payers. Moreover, after a product is approved or labeled, it is generally too late to conduct informative comparative research.

Current status: The FDA

The FDA plays a gatekeeper role to ensure that poor quality information is not used as the basis of approval or labeling. Historically, many patient questionnaires were not well developed and generated untrustworthy data. The FDA produced a PRO Guidance (draft 2006; final 2009) that was a major advancement toward establishing methodological standards for developing and using PRO measures [2]. But the science of PRO measurement and the community of experts in this field have advanced substantially over the past decade. There have been critiques from some members of this community that overly stringent application of Guidance principles by the FDA has hindered rather than promoted inclusion of PROs in labels. The article by DeMuro et al. [1] in this issue of *Value in Health* reports that 25% of labels since 2006 include PRO end points. It is debatable whether this represents a triumph or failure of the Guidance or of the movement toward making drug labels more patient-centered. It can be spun either way, although the article's authors suggest that 25% is a small number given the 50% of drug approval packages that include PRO end points. But the more salient questions here are whether the FDA is appropriately critiquing current uses of PROs, whether it is feasible for sponsors to meet FDA standards in most cases, and whether the right PROs are being integrated into the right studies. The article substantiates what many sponsors and the FDA have anecdotally pointed out: that there is still quite a bit of heterogeneity in how PRO end points are designed by sponsors and in how they are considered across and within FDA review divisions.

Moving forward: The sponsor role

So where are we to go from here? Below (and summarized in Table 1) is a proposed path forward for sponsors:

Every drug development program should consider early on how information elicited from patients could be informative to decision makers who will ultimately use, prescribe, or pay for a product. This includes assessment of symptoms that may be alleviated by a product, symptomatic side effects, and changes in overall HRQOL or health state. An underlying fundamental change in culture is necessary, in which the value and feasibility of including PROs is understood by clinical investigators and leadership. Operationalization involves the following:

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

Table 1 – Recommendations for industry sponsors and for the FDA to consider toward increasing success including patient-reported outcome (PRO) end points in pivotal trials and US drug labels, without compromising methodological rigor.

Recommendations for sponsors	Recommendations for the FDA
1. Create ongoing relationships between clinical development teams and internal or external PRO experts.	1. Increase internal FDA PRO measurement expertise, both within review divisions and by expanding SEALD.
2. Evaluate potential value of PROs in every clinical development program, starting during early-phase research. Consider how PROs would inform decision makers including patients, clinicians, regulators, and payers.	2. Consider PROs as essential information to understand the properties of a product, without which a submission is incomplete.
3. Engage FDA early to discuss the role of PRO end points and specific measurement strategies.	3. Encourage sponsors to incorporate and develop PRO measures early in a product development program.
4. Conduct early qualitative research to identify outcomes important to patients including symptoms of disease, symptomatic toxicities of a product, and impact of the product on global health-related quality of life.	4. Relax stringency around accepting health-related quality-of-life data for inclusion in labels.
5. Include PROs as primary or secondary end points in pivotal trials, with adequate statistical power. Provide a rationale if PROs are not included as an endpoint (such as if the impact of the product on the patient experience is already known).	5. Adjust criteria for concluding an established PRO measure is fit for purpose in a new target population or context, to require only limited qualitative but not further quantitative evidence.

FDA, Food and Drug Administration; SEALD, Study End points and Labeling.

- As soon as the activity of a product in a particular target population is identified, work should begin to see what symptoms are important in patients representing that population (unless already well known). A small number of one-on-one interviews, focus groups, or multisymptom screening surveys in untreated patients plus a literature review can quickly identify whether there are specific symptoms or functional impairments of importance. Early information can be gathered to explore whether the treatment alleviates any of the baseline symptoms and whether there are symptomatic toxicities associated with the treatment. These become the PROs of interest.
- Next, measures must be identified or developed to assess these PROs. Again, the earlier the better in a program. This is where the FDA Guidance has raised the bar a bit, but it is not insurmountable by any means. The most important step is to ensure that measures being used are considered meaningful in the target population, which can be done through interviews in a small number of patients [3]. Translations should be considered early if an ultimate multinational study is foreseen, and a number of companies specialize in translating PRO tools efficiently and inexpensively in keeping with established standards [4]. Early engagement of FDA reviewers to ensure that a plan is consistent with FDA expectations is highly advisable.
- PRO measures can be particularly informative in dose finding and should be considered for use to ensure that there are not excessive side effects from the patient perspective—an element that is frequently ignored in drug development, leading to potential selection of inappropriately high doses (which real-world patients may ultimately not wish to endure).
- PRO measures are useful and should be considered in phase 3 trials to 1) demonstrate comparative benefits or comparative tolerability from the patient perspective; 2) enhance progression-free survival end points and surrogate end points to substantiate that a product impacts how a patient feels or functions; and 3) screen for symptomatic adverse events from the patient perspective (notably, patients better detect baseline symptoms than staff, and so when symptoms are detected during a study it is more clear via PROs whether they were preexisting) [5]. Again, it is as useful to decision makers to know that there is no effect on symptoms as to know that there is, and so PRO data are always informative. PRO end points should be included as primary or secondary end points with adequate statistical power reserved for them. Measures to minimize missing data, such as backup data collection methods and reminders to patients, should be employed, as well as an *a priori* plan for imputing missing PRO data.

Moving forward: The FDA role

Below (and summarized in Table 1) is a proposed path forward for the FDA:

FDA reviewers should consider the potential role of PROs to support understanding of the properties of *every* product and should consider a sponsor's research plan to be incomplete if the direct patient perspective is not represented (or if a justification for not collecting the patient perspective is not included). Operationalization involves the following:

- The FDA needs more expertise on PRO measurement. Widespread interest in a more patient-centered regulatory process and the heterogeneity in approaches identified in the new article in this issue of *Value in Health* indicates that this is an acute need. To get there necessitates a three-pronged approach: 1) reviewers must become better versed in the methods of PRO measurement; 2) statisticians with expertise in analyzing PRO data must be developed, hired, or contracted; and 3) the staff of the FDA's internal PRO resource, Study Endpoints and Labeling (SEALD), must be expanded with well-qualified and thoughtful individuals who communicate effectively with the review divisions and who have sufficient experience in PRO measure development and clinical research to be realistic about the balance between the rigor and feasibility of implementing PRO end points. At least six hires of mid-level professional staff or contractors into SEALD with associated administrators is necessary to address existing needs. Recent FDA requests to use congressional allocations to support PRO expertise could be directed in these three areas.
- FDA reviewers should engage sponsors early to emphasize the importance of including PROs in development research when the impact of a product on the patient experience is not already known (or require a justification for why it is not included). Applications and proposed drug labels without information about the patient subjective experience should be considered incomplete. Abundant research demonstrates that no other source of information can substitute for patient direct reports, and information about symptoms from other sources such as clinicians substantially underestimates prevalence and sever-

ity. Reviewers should assist sponsors to determine methods acceptable to FDA for collecting this information.

- The FDA should consider relaxing its stringency about generic or HRQOL tools. These should be viewed as acceptable for labeling purposes if demonstrated to have robust measurement properties. An outdated argument against such measures has been that it is unclear what exactly they are measuring and that they represent a composite of experiences including symptomatic improvement, toxicity, functional status, or psychosocial status. But what is more important than quality of life? Arguably, most other regulatory end points also represent multifactor common final pathways (including overall survival). Single items asking patients about their quality of life perform well from a methodological standpoint and are highly correlated with meaningful outcomes such as symptoms, performance status, disease regression/progression, and survival. There is a compelling case for these end points to be accepted as a basis for labeling (or as supportive of specific symptom end points) as they are the most meaningful and important to patients in many cases. Moreover, generic measures are useful in economic and comparative effectiveness analyses, which are standard in Europe and increasingly important in the United States.
- The FDA should consider adjusting its criteria for fitness for purpose of PRO measures. It could be regarded as sufficient in most cases for a sponsor to use an existing measure with demonstrated good measurement properties in another population, with qualitative research in the new target population showing that items are meaningful and understood. This could be an acceptable criterion for concluding fitness for purpose. Requiring new validation or establishment of clinically meaningful score changes is probably not necessary in more than a couple of different populations.

Needed regulatory science methods research in PROs

There are several known methodological knowledge gaps that present barriers to PRO end points being accepted for inclusion in US labels, as underlined by the article in this issue of *Value in Health*. Improving methods and knowledge in these areas will allow sponsors and regulators to feel more comfortable about the fidelity of PRO analyses:

- Use of PROs in open-label studies or in studies with inadvertent unbinding of treatment allocation: It is theorized that patient self-reports are biased when patients believe they are receiving an active or superior treatment (e.g., a patient realizes he or she is on the experimental arm of a study, and this leads him or her to report greater pain improvement). Published literature reports variable effects of such bias on patient reporting, but this concept serves as an underlying basis for blinding in trials. In general, the FDA will not accept PRO endpoint data from trials that are open-label or that are difficult to blind because of typical or observable side effects associated with one of the treatments (e.g., rash with tyrosine kinase inhibitors). The magnitude of this potential bias is not known. It has been suggested that a sufficiently large effect size requirement could overcome this source of bias, if it exists. Research in this area is warranted to inform study design and review.
- Approach to missing PRO data: Rates of missing data are highly variable between trials. Research is needed to identify effective approaches to minimizing missing data including backup data collection techniques. In addition, standard approaches for imputing missing patient-reported data are needed.
- Approach to PRO data in multinational or multicultural trials: Patient responses to PRO questionnaires can vary on the basis of cultural differences in perceptions of the domains of interest. For example, beliefs about pain and pain management may differ between cultures. This does not negate the value of using a measure in a trial spanning cultures, but it may be important to have a balanced number of patients between arms in major subcultures. Research on how to accommodate for these differences in study results would improve confidence in using PRO end points in large multinational trials.

Over the past 10 to 15 years, there has been substantial progress in the methodological science and technical feasibility of collecting data directly from patients. An increasing general interest in patient-centeredness, and recognition that the patient perspective is currently underrepresented in pivotal trials and in US drug labels, suggests a need for a change in orientation and operationalization by both industry sponsors and FDA reviewers. The FDA's PRO Guidance was a major step forward, and there is now evidence that almost 25% of US drug labels include PRO end points. But much progress remains. PRO end points are clearly appropriate in many more than these labels, and even in the existing labels with PRO end points, the picture of the patient experience is not comprehensive in many cases. Systematic processes both in companies and in the FDA to assess early on what outcomes are important to patients, and how they should be measured, should become *de rigeur*. Until then, we will be left with an incomplete picture of how products impact end users, and patients will remain unclear on how "patients like them" felt when using products.

Ethan Basch, MD, MSc
Department of Medicine and Health Outcomes Research Group, Memorial Sloan-Kettering Cancer Center, New York, NY, USA

1098-3015/$36.00 – see front matter

Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.jval.2012.03.1385

REFERENCES

[1] DeMuro C, Clark M, Mordin M, et al. A review of patient-reported outcomes labels in the US: 2006–2010. Value Health 2013;15:437–42.
[2] U.S. Department of Health and Human Services, Food and Drug Administration. Guidance for industry: patient-reported outcomes measures: use in medical product development to support labeling claims. Issued December 2009. Available from: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM19328.pdf. [Accessed March 24, 2010].
[3] Wild D, Eremenco S, Mear I, et al. Multinational trials-recommendations on the translations required, approaches to using the same language in different countries, and the approaches to support pooling the data: the ISPOR Patient-Reported Outcomes Translation and Linguistic Validation Good Research Practices Task Force Report. Value Health 2009;12:430–40.
[4] Patrick DL, Burke LB, Gwaltney CJ, et al. Content validity–establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO Good Research Practices Task Force report: part 2–assessing respondent understanding. Value Health 2011;14:978–88.
[5] Basch E. The missing voice of patients in drug-safety reporting. N Engl J Med 2010;362:865–9.

VALUE IN HEALTH 15 (2012) 437–442

Available online at www.sciencedirect.com

SciVerse ScienceDirect

journal homepage: www.elsevier.com/locate/jval

Patient-Reported Outcomes

A Review of Patient-Reported Outcome Labels in the United States: 2006 to 2010

Ari Gnanasakthy, MSc[1,], Margaret Mordin, MS[2], Marci Clark, PharmD[2], Carla DeMuro, MS[2], Sheri Fehnel, PhD[2], Catherine Copley-Merriman, MS[2]*

[1]*Novartis Pharmaceuticals Corporation, East Hannover, NJ, USA;* [2]*RTI Health Solutions, Durham, NC, USA*

A B S T R A C T

Objective: In 2004, Willke and colleagues reviewed the efficacy endpoints reported in the labels of new drugs approved in the United States from 1997 through 2002 to evaluate the use of patient-reported outcome (PRO) endpoints. Of the labels reviewed, 30% included PROs. Our study aimed to build on this work by describing the current state of PRO label claims granted for new molecular entities (and biologic license applications since February 2006 after the release of the US Food and Drug Administration (FDA) draft PRO guidance. **Methods:** All new molecular entities and biologic license applications approved by the FDA from January 2006 through December 2010 were identified by using the Web page of the FDA Drug Approval Reports. For all identified products, drug approval packages and approved product labels were reviewed to identify PRO endpoint status and to determine the number and type of PRO claims. **Results:** Of the 116 products identified, 28 (24%) were granted PRO claims; 24 (86%) were for symptoms, and, of these, 9 (38%) claims were pain related. Of the 28 products with PRO claims, a PRO was a primary endpoint for 20 (71%), all symptom related. **Conclusions:** The FDA continues to approve PRO claims, with 24% of new molecular entities and biologic license applications being granted. Successful PRO label claims over the past 5 years have generally supported treatment benefit for symptoms specified as primary endpoints.

Keywords: drug labeling, patient-reported outcomes.

Introduction

The content of package inserts from the US Food and Drug Administration (FDA) is vital to the commercial success of a medicinal product. These package inserts, also called product labels, constitute the formal, government-approved definition of a drug's benefits and risks. Package inserts are written by (and are the property of) the manufacturer but require FDA approval; they define the boundaries of the legal promotion of a product's properties [1].

Patient-reported outcome (PRO) is an umbrella term that comprises a range of potential measurement endpoints, but it is used specifically to describe outcomes collected directly from the patient, without interpretation by clinicians or others [2,3]. PRO use is particularly common for products developed to treat chronic, disabling conditions where the intention is not necessarily to cure but to ameliorate symptoms, facilitate functioning, or improve quality of life. PROs are the primary endpoints in clinical trials evaluating drug products for disease areas such as irritable bowel syndrome, migraine, and pain. PROs provide key supportive data in many other disease areas, such as insomnia, asthma, and psychiatric disorders. In oncology, PROs are commonly used to assess both treatment benefits and toxicity to fully evaluate the impact of treatment on health-related quality of life (HRQOL). PROs can also be used in clinical trials to assess treatment satisfaction, compliance, and caregiver burden.

Increase in the use of formal questionnaires in clinical trials [4], advances in methodological rigor in measurement science during the 1980s and the 1990s [5], and the need to standardize the terminology [2] led to the guidance on PROs from the FDA, especially because it is related to drug labeling and promotion.

For drug manufacturers seeking PRO claims, the FDA's release of a draft guidance in 2006 [6] and a final guidance in 2009 [7] (*Guidance for Industry. Patient Reported Outcomes: Use in Medical Product Development to Support Labeling Claims* [PRO guidance]) was a landmark event. The PRO guidance describes the use of PROs to support potential claims in product labeling. Based on this PRO guidance document, PROs may be used to support treatment benefit claims in FDA-approved product labeling. The claims must be supported by appropriately designed investigations using PROs that have been demonstrated to measure the concept underlying the claim [3].

International societies have held workshops to debate the impact of the FDA PRO guidance, and journals have hosted special issues devoted entirely to this topic [8]. It is generally agreed that the PRO guidance has set a high standard for developing and implementing PRO measures in clinical trials for new drug products and has also provided a blueprint for sponsors who wish to obtain PRO label claims for their products [9].

A review of PRO labels granted from 1997 through 2002 [10] showed that PRO evidence was cited in the Clinical Studies section of

* *Address correspondence to:* Ari Gnanasakthy, Novartis Pharmaceuticals Corporation, One Health Plaza, East Hanover, NJ 07936, USA.
E-mail: ari.gnanasakthy@novartis.com.

1098-3015/$36.00 – see front matter

doi:10.1016/j.jval.2011.11.032

the label for 30% of the new product approvals and 11% of the new products were approved on the basis of PROs alone. Our study aimed to build on the work by Willke et al. [10] by describing the current state of PRO label claims granted for new molecular entities (NMEs) and biologic license applications (BLAs) from 2006 through 2010.

The purpose of this study was to compile and review the PRO label claims granted over the 5 years since the release of the draft PRO guidance (2006–2010) [6]. Moreover, we were interested in understanding the types of claims granted based on PROs. We hypothesized that PRO claims would be more likely for first-order impact assessments such as symptoms, rather than for more complex concepts such as HRQOL.

Methods

Products reviewed in this analysis included new drugs approved in the United States from January 2006 through December 2010. The Web page of the FDA Drug Approval Reports was used to determine the number of products approved in the time period of interest. The report options selected were original new drug approvals (NDAs) and BLAs by month; months were selected sequentially beginning with January 2006 and ending in December 2010. The reports include a specification of the Center for Drug Evaluation Research NDA chemical classification. Our review included products classified by the Center for Drug Evaluation Research as NMEs or BLAs. Therefore, we excluded products containing substances previously marketed with a different brand name or a set of indications, as a different dosage form or strength, or as a combination product of previously marketed entities.

Once products were identified, drug approval packages (DAPs) and approved product labels were reviewed. As available, information was retrieved from the Medical Review, Summary Review, Cross-Discipline Team Leader Review, and other review sections from the DAP, as well as the Indication and Clinical Studies section of the approved product label. The DAPs were located on the FDA's Web site Drugs@FDA (www.accessdata.fda.gov). In most cases, the product label was also found on this FDA Web site under approval history. In the event the approved label was unavailable for the specified time frame, the current label was evaluated. As available, the following information was collected for each US drug product identified:

- Brand name
- Generic name
- Date of approval
- Applicant
- Label indication
- PRO claim language
- PRO measures named in label
- Reviewing division
- Medical review available (yes/no)
- Indication in DAP of Study Endpoints and Label Development (SEALD) review (yes/no) and comments
- PRO measures mentioned in the label and DAP, and endpoint status (primary, secondary, tertiary/exploratory)
- PRO results reported as statistically significant (yes/no)

PRO claim language from the Indication and Clinical Studies sections of the label was reviewed and characterized as symptoms (yes/no), functioning (yes/no), HRQOL (yes/no), patient global rating (PGR) (yes/no), or other (yes/no). A single rater applied standard definitions to the review of the labels for characterization.

Symptoms were defined as any subjective evidence of a disease, health condition, or treatment-related effect that can be noticed and known only by the patient. Functioning claims related to restriction or lack of ability to perform an activity(ies) in the manner or within the range considered normal. HRQOL claims were defined as those referencing a multidomain concept representing the patient's general perception of the effect of illness and treatment on physical, psychological, and social aspects of life. A PGR was defined as any assessment or evaluation of the patient's disease or condition identified as "global." These classifications are in line with the definitions provided in the final PRO guidance and the work previously reported by Caron et al. [11]. A product label may contain more than one PRO claim.

Statistical analysis consisted of frequencies and cross tabulations of measured characteristics. Calculations were performed by using Microsoft Excel 2007.

Results

A total of 156 new drugs were approved during this period. DAPs were located and reviewed for all 156 products. The DAPs for all products contained medical reviews. Some DAPs also included summary reviews or cross team leader reviews or both. Product labels were located for all the products. Of the 156 approvals, 33 were granted tentative approvals and full approval will not be granted until after patent exclusivity expires. Because these were for generic products, we excluded them from our analysis. Some drugs approved during this period were subsequently removed from the market but were nonetheless included in this analysis. Denosumab, although registered as both Prolia and Xgeva, was considered a single new drug with the same clinical studies for registration and a single BLA supporting both. Similarly, Natazia was considered a single new drug because the same clinical studies were used in the registration files and a single NDA supported it. Sabril was considered a single new drug despite two unique NDA numbers supporting the different formulations. Finally, there were four new products approved with no data available on the FDA's Web site, including a label, at the time of our data extraction and analysis and so these products were excluded from this review. Therefore, a total of 116 products were included in this review.

Of the 116 products reviewed, the largest number (n = 16) was reviewed by the Drug Oncology division, followed by Neurology Products (n = 11), Cardiovascular and Renal Products (n = 10), and Anesthesia, Analgesia, and Rheumatology Products (n = 10) divisions (Table 1). PRO claims appeared in 28 product labels (24% of the 116 products reviewed) across 11 reviewing divisions. The indications for all 28 products with PRO claims in the label are available in Appendix A in Supplemental Materials found at doi: 10.1016/j.jval.2011.11.032). Among the 28 products with PRO claims in the label, Neurology Products (n = 7; 25.0%) and Anesthesia, Analgesia, and Rheumatology Products (n = 6; 21.4%) divisions granted the most PRO label claims. Approximately two-thirds of the products reviewed by the Neurology Products; Anesthesia, Analgesia, and Rheumatology Products; and Pulmonary and Allergy Products reviewing divisions received PRO claims in the label. The following reviewing divisions did not grant any PRO label claims: Drug Oncology; Biologic Oncology; Antiviral, Dermatology, and Dental Products; and Special Pathogen and Transplant Products.

The 28 products received a total of 38 PRO label claims (Table 2). The majority of the products (n = 20; 71%) received one PRO label claim. The products with one PRO label claim were characterized as follows: symptoms (n = 16), functioning (n = 1), HRQOL (n = 1), and other (n = 2). Of the eight products that received multiple PRO label claims, six received two PRO label claims and two products received three PRO label claims (symptoms, functioning, and PGR). Of the 28 products with PRO label claims in the Clinical Studies section of the label, 14 (50%) also contained PRO claims within their indication statements. Only one of the claims reviewed appeared in the medication guide. There were no PRO claims related to decrements in health.

Most PRO label claims granted were for symptoms (85.7%) and functioning (25%) (Table 2). A few products (n = 3; 10.7%) received PRO label claims on the basis of PGRs (e.g., seizure severity and

Table 1 – Number of products approved and number of PRO claims granted by reviewing divisions.

Reviewing division	Products reviewed	Number of products approved	Number of products that include a PRO claim
Anesthesia, Analgesia and Rheumatology Products	Chantix,* Arcalyst,* Nucynta,* Lusedra, Savella,* Uloric, Simponi,* Ilaris, Actemra,* Xiaflex	10	6
Antimicrobial Products	Durezol*	1	1
Anti-infective and Ophthalmology Products	Lucentis, Altabax, Doribax, Besivance, Vibativ, Bepreve,* Lastacaft,* Teflaro	8	2
Antiviral Products	Prezista, Tyzeka, Selzentry, Isentress, Intelence, PegIntron/Rebetol Combo Pack, acyclovir, hydrocortisone, Zidovudine	8	0
Biologic Oncology Products	Vectibix, Arzerra	2	0
Cardiovascular and Renal Products	Tekturna, Letairis,* Bystolic, Cleviprex, Samsca, Tyvaso, Effient, Multaq, Asclera,* Pradaxa	10	2
Dermatology and Dental Products	Veregen, Ulesfia, Stelara	3	0
Drug Oncology Products	Dacogen, Sprycel, Zolinza, Tykerb, Torisel, Ixempra Kit, Tasigna, Treanda, Firmagon, Mozobil, Afinitor, Folotyn, Votrient, Istodax, Jevtana, Halaven	16	0
Gastroenterology Products	Myozyme, Elaprase, Cimzia,* Relistor, Entereg, Vpriv, Carbaglu, Lumizyme	8	1
Medical Imaging and Hematology Products	Soliris,* Ammonia N 13, Mircera, Lexiscan, Eovist, Nplate, AdreView, Promacta, Ablavar	9	1
Metabolism and Endocrinology Products	Januvia, Somatuline Depot, Kuvan, Onglyza, Livalo, Victoza, Egrifta*	7	1
Neurology Products	Azilect,* Neupro, Xenazine, Vimpat,* Banzel,* Dysport,* Extavia, Sabril 500-mg tablet,* Ampyra,* Xeomin,* Gilenya	11	7
Nonprescription Clinical Evaluation Products	Anthelios SX, Cetirizine Hydrochloride Allergy,* Cetirizine Hydrochloride Hives Relief*	3	2
Psychiatry Products	Invega , Vyvanse,* Pristiq, Fanapt, Invega Sustenna, Saphris, Latuda	7	1
Pulmonary and Allergy Products	Omnaris,* Kalbitor,* Krystexxa	3	2
Reproductive and Urologic Products	Toviaz,* Rapaflo,* Natazia, Ella, Prolia	5	2
Special Pathogen and Transplant Products	Eraxis, Noxafil, Pylera, Coartem, Zortress	5	0
Total		116	28

PRO, patient-reported outcome.
* Products with PRO claims in the label.

global impression of change). Two products were granted PRO claims classified as other: these were patient satisfaction with treatment (Asclera) and distress associated with belly appearance (Egrifta). Pain continues to be a prominent symptom among the PRO label claims granted, ranking highest (n = 7) among the 16 symptoms label claims followed by allergy-related symptoms (n = 5). The concepts of pain and reduced pain appear straightforward, and as such, little was discussed in the DAPs regarding the measurement of pain itself. Pain assessments via visual analogue scales and numeric rating scales are common, with little (if any) discussion in the DAPs surrounding the question stem or anchors used.

Table 2 – Types of claims granted.

Type of claim	All products with PRO claims (N = 28)		Pain products excluded (N = 21)	
	n	%	n	%
Symptoms	24	85.7	14	66.7
Functioning	7	25.0	3	14.3
HRQOL	2	7.1	2	9.5
PGR	3	10.7	1	4.8
Other	2	7.1	2	9.5

HRQOL, health-related quality of life; PGR, patient global rating.

More than 30 different PRO measures were used to support the PRO claims received (Table 3). The bulk of the measures were designed to measure a single concept such as pain or seizure rates (n = 8) or diary assessments (n = 6). Another large proportion of the measures appears to be expected by the reviewing divisions given their familiarity with the measures (n = 9) (e.g., Health Assessment Questionnaire, Short Form 36 Health Survey, and International Prostate Symptom Score). We noted several hybrid measures that combined both clinician-reported outcomes and PROs into a single measurement tool (e.g., Toronto Western Spasmodic Torticollis Rating Scale and Activities of Daily Living and Motor subscales of the Unified Parkinson's Disease Rating Scale). Although these hybrid measures are not solely patient reported, they contain PROs that are critical to assessing efficacy in the given indications.

The extent of information identified in the label regarding the specific PRO measures used to support the label claim was variable. Some labels included very little information regarding the PRO assessment. For example, the assessment of ocular itching in the Lastacaft label is not described at all ("more effective than its vehicle in preventing ocular itching in patients with allergic conjunctivitis induced by ocular allergen challenge"). Other labels included more specific information regarding the PRO assessment. For instance, the Egrifta label has a subsection on PRO within the Clinical Studies section. The description of patient-rated degree of distress in the Egrifta label includes the following:

Table 3 – Measures used to support PRO label claims.

Type of claim/product	Measure description supporting claims
Symptoms	
Azilect	Diary: "On/off" periods
Chantix	Brief Questionnaire of Smoking Urges and Minnesota Nicotine Withdrawal Scale
Omnaris	Diary: Nasal symptoms (runny nose, nasal itching, sneezing, and nasal congestion)
Vyvanse	Conners' Parent Rating Scale
Soliris	Functional Assessment of Chronic Illness Therapy—Fatigue*
Arcalyst	Diary: Signs and symptoms of cryopyrin-associated periodic syndrome: joint pain, rash, feeling of fever/chills, eye redness/pain, and fatigue
Cimzia	Crohn's Disease Activity Index†
Durezol	Visual analogue scale—eye pain/discomfort*
Toviaz	Diary: Urge urinary incontinence episodes and number of micturitions (frequency)*
Rapaflo	International Prostate Symptom Score
Vimpat	Seizure frequency
Banzel	Seizure severity from the Parent/Guardian Global Evaluation of the patient's condition
Nucynta	Pain numeric rating scale
Savella	Pain visual analogue scale
Dysport	TWSTRS†
Simponi	Health Assessment Questionnaire—Disability Index and Bath Ankylosing Spondylitis Functional Index†
Cetirizine hydrochloride-allergy	Diary: Symptoms include sneezing, rhinorrhea, nasal pruritus, ocular pruritus, tearing, and redness of the eyes
Cetirizine hydrochloride-hives	Diary: Severity and duration of hives and pruritus
Sabril	Complex partial seizures—seizure frequency
Bepreve	Ocular itching
Kalbitor	Mean Symptom Complex Severity and Treatment Outcome Score
Actemra	Pain visual analogue scale
Xeomin	TWSTRS subscales†
Lastacaft	Ocular itching
Function	
Azilect	Activities of Daily Living and Motor subscale of the Unified Parkinson's Disease Rating Scale†
Savella	Physical function (Short Form 36 Health Survey physical component summary)
Dysport	TWSTRS subscales†
Simponi	RA and PSA: Health Assessment Questionnaire—Disability Index and AnkSpon: Bath Ankylosing Spondylitis Functional Index†
Ampyra	12-Item Multiple Sclerosis Walking Scale
Actemra	Health Assessment Questionnaire
Xeomin	TWSTRS subscales†
HRQOL	
Soliris	European Organisation for the Research and Treatment of Cancer—Quality of Life Questionnaire Core 30 Items*
Letairis	Short Form 36 Health Survey
PGR	
Banzel	Seizure severity from the Parent/Guardian Global Evaluation of the patient's condition
Savella	Patient global assessment of change
Simponi	Patient global assessment of change
Other	
Asclera	Patient satisfaction (verbal rating scale)
Egrifta	Distress associated with belly appearance

HRQOL, health-related quality of life; PGR, patient global rating; PRO, patient-reported outcome; TWSTRS, Toronto Western Spasmodic Torticollis Rating Scale.

* Not mentioned in label.

† Hybrid clinician-reported and patient-reported measure.

Patients rated the degree of distress associated with their belly appearance on a 9-point rating scale that was then transformed to a score from 0 [extremely upsetting and distressing] to 100 [extremely encouraging]. A score of 50 indicated neutral [no feeling either way]. A positive change from baseline score indicated improvement, i.e., less distress.

In addition, certain therapeutic areas included extensive information regarding the PRO assessment, where it was the primary efficacy endpoint (e.g., Omnaris label for seasonal allergic rhinitis).

A PRO was the primary endpoint for 20 of the 28 (71%) products with PRO label claims (Table 4). All 20 primary endpoints were based on symptoms. PRO label claims were granted for nonprimary endpoints for 8 of the 28 (29%) products. The four products for which a PRO was not a primary endpoint and where a symptom claim was not granted were those granted PRO claims for distress (Egrifta), satisfaction (Asclera), HRQOL (Letairis), and functioning (Ampyra).

Three products received PRO claims on the basis of PGRs. These included a measure of seizure severity from the parent/guardian global evaluation of the patient's condition (Banzel), a patient's global assessment of disease activity (Simponi), and a patient global impression of change (Savella). The Banzel label specifies the PGR as one of the three primary efficacy variables:

Table 4 – PRO primary endpoint and symptom claims.

	PRO primary endpoint		Total number of products
	Yes (n = 20)	No (n = 8)	
Symptoms claim: Yes	20	4	24
Symptoms claim: No	0	4	4
Total	20	8	28

PRO, patient-reported outcome.

> . . . seizure severity from the Parent/Guardian Global Evaluation of the patient's condition. This was a 7-point assessment performed at the end of the Double-blind Phase. A score of +3 indicated that the patient's seizure severity was very much improved, a score of 0 that the seizure severity was unchanged, and a score of 3 that the seizure severity was very much worse.

Results of the three primary endpoints, including the PGR, are presented within a table within the label.

Within the DAPs, SEALD was mentioned as providing a review for the following four products: Chantix, Soliris, Cimzia, and Egrifta. In the case of Chantix, SEALD personnel were named and their specific comments were directly available for review as part of the DAP, whereas for Soliris and Egrifta, a summary of the SEALD review and comments was referenced in the context of the medical team's review but specific comments from identified SEALD reviewers were not available. For Cimzia, only the names of FDA personnel involved with the product review revealed SEALD team involvement. There was no evidence in the other DAPs as to whether SEALD provided additional consultation to the reviewing division.

The extent of publicly available information regarding the labeling discussion itself is limited because proposed labeling language is considered proprietary. Nevertheless, the Egrifta DAP reveals that in response to the comments and recommendations made by the SEALD consult, the clinical and statistical team decided to include in the label only the results of the belly appearance distress. It was noted by the Reviewing Division that

> from a clinical perspective Belly Appearance Distress is an endpoint of higher significance as it does not measure the self-reported perception about changes in the size of the abdomen but rather the emotional impact and distress for the patient, an important proxy for QOL in HIV-patients with lipodystrophy. As recommended by the SEALD consult, the term XXX will no longer be included in Egrifta label description of this PRO since, although developed with advice, it no longer meets the new standard set by the December 2009 FDA PRO guidance.

Discussion

This review provides a compilation and categorization of PRO label claims granted since the release of the draft PRO guidance in 2006 [6]. Although a similar review of PRO labels for the 3 years immediately before the release of the draft PRO guidance is not available (2003–2005), this review provides an opportunity to compare the current state of PRO label claims over the 5-year period since the release of the draft PRO guidance (2006–2010) with that reported by Willke et al. between 1997 and 2002 [10].

Despite the hope that after the release of the PRO guidance the proportion of NMEs and BLAs with PRO label claims would increase given the established guidelines [12], our results indicate that this proportion has decreased slightly from 30% reported by Willke et al. to 24%. Our findings, based on NMEs and BLAs, are similar to the 21.5% reported by Marquis et al. [13], which was based on all products over the same time period as of our review.

The guidance from the FDA has provided the pharmaceutical industry with much more information regarding regulatory expectations than ever previously available. Our review, however, suggests that there is disparity across reviewing divisions in terms of the proportion of PRO label claims granted. Although several reviewing divisions have granted PRO label claims, others have yet to grant any since the release of the draft PRO guidance in 2006.

For example, although the FDA guidance for industry on oncology clinical trial design cites symptoms as a direct efficacy endpoint that can be used to support regulatory approval [14,15], only 4 of the 57 approvals from 1990 through 2002 were based on decrease in tumor-specific symptoms [16]. Gondek et al. [14] reported findings of an analysis of PRO claims among product labels for oncology. From a pool of 70 new or revised product labels (from January 2002 through September 30, 2006), there were six labels for a new product or a new indication that contained PRO assessments based on symptoms (n = 5) and functions (n = 1) [14]. Yet there have not been any PRO label claims for oncology products since the release of the draft PRO guidance.

Occurrences of symptoms are the most commonly reported PRO label claims granted. This finding, based on the analysis of NMEs and BLAs since the release of the draft PRO guidance in 2006 [6], is similar to the findings from previous analyses of all PRO labels in the United States [17] and Europe [11]. The dominance of symptom-based PRO claims may be twofold. First, symptoms are typically the first-order impact of many diseases and treatments. Second, most symptom occurrences that can be quantified by frequency, severity, and duration are easy to measure on simple scales and with patient diaries in clinical trials conducted in multiple regions.

Our analyses also show that many PRO measures used for the purpose of label claims can be considered to be well established in the literature on the basis of the frequency of use in clinical trials and available information on their development and psychometric measurement properties (e.g., Short Form 36 Health Survey and Health Assessment Questionnaire).

Patient diaries continue to be used prolifically in capturing PRO data. Diaries capture simple items, such as seizure frequency, severity and duration of pruritus, and the on/off cycle of Parkinson's symptoms, and thus tend to result in simple PRO labeling claims (e.g., reduction in 28-day seizure frequency and prevention of itching).

PRO label claims for nonprimary endpoints are uncommon from the FDA. There are three likely reasons for this.

1. Primary and nonprimary endpoints tend to measure the same domain. Furthermore, such claims, overemphasizing the efficacy of products, are often the target for warning letters from the Division of Drug Marketing, Advertising, and Communications [18].
2. Sponsors are unlikely to commit resources for nonprimary endpoints during the early stages of product development, which can be substantial for developing new PRO measures aimed at multinational studies, when the likelihood of changes to the target product profile and the rate of attrition are still high.
3. Sponsors may be reluctant to support the logistical complexities related to nonprimary PRO endpoints during the execution of a multinational study. For example, protocol amendments, such as changes to inclusion and exclusion criteria for patient characteristics while the study is ongoing, will preserve the integrity of the primary endpoint but may affect the validity of the nonprimary PRO endpoint. In addition, slow patient recruitment in studies may necessitate the need to close study centers in some countries and open new centers in other countries. Sponsors are unlikely to wait for the availability of new translations of PRO measures and take time to implement data col-

lection logistics, which may delay the study by months, to support nonprimary PRO endpoints.

This review was based on information publicly available in DAPs on FDA's Web site. Additional material, of which we were unaware or that was unavailable to us, may have been considered as part of the FDA review. Furthermore, SEALD acts on a consultancy basis and therefore not all reviews received (or required) their input regarding the PROs.

Conclusions

The percentage of NMEs and BLAs with PRO label claims has decreased from 30% [10] reported between 1997 and 2002 to 24% between 2006 and 2010. PRO label claims are granted mostly for primary endpoints that are also symptoms. The majority of accepted claims are supported by simple scales, such as a visual analogue scale, a numeric rating scale, or symptom diaries, or on the basis of measures that have been traditionally accepted by the reviewing divisions. Examination of future sponsor submissions and regulatory feedback for studies planned and executed since the release of the final PRO guidance may provide additional insight into how to increase success in obtaining PRO-based label claims.

Acknowledgments

We gratefully acknowledge the research assistance of Emily Evans in the development of this article. We also gratefully acknowledge Lynda Doward for her review of the article.

Source of financial support: This study was funded by Novartis Pharmaceuticals Corporation.

Supplemental Materials

Supplemental material accompanying this article can be found in the online version as a hyperlink at doi:10.1016/j.jval.2011.11.032 or, if a hard copy of article, at www.valueinhealthjournal.com/issues (select volume, issue, and article).

REFERENCES

[1] Avorn J, Shrank W. Highlights and a hidden hazard: the FDA's new labeling regulations. N Engl J Med 2006;354:2409–11.
[2] Acquadro C, Berzon R, Dubois D, et al. Incorporating the patient's perspective into drug development and communication: an ad hoc task force report of the Patient-Reported Outcomes (PRO) Harmonization Group meeting at the Food and Drug Administration, February 16, 2001. Value Health 2003;6:522–31.
[3] Burke LB, Kennedy DL, Miskala PH, et al. The use of patient-reported outcome measures in the evaluation of medical products for regulatory approval. Clin Pharmacol Ther 2008;84:281–3.
[4] Shah SN, Sesti AM, Copley-Merriman K, et al. Quality of life terminology included in package inserts for US approved medications. Qual Life Res 2003;12:1107–17.
[5] Wood-Dauphinee S. Assessing quality of life in clinical research: from where have we come and where are we going? J Clin Epidemiol 1999; 52:355–63.
[6] US Department of Health and Human Services. Draft guidance for industry. Patient-reported outcome measures: use in medical product development to support labeling claims. February 2006. Available from: http://www.fda.gov/ohrms/dockets/98fr/06d-0044-gdl0001.pdf. [Accessed January 14, 2011].
[7] US Department of Health and Human Services. Guidance for industry. Patient-reported outcome measures: use in medical product development to support labeling claims. December 2009. Available from: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM193282.pdf. [Accessed January 14, 2011].
[8] Sloan JA, Halyard MY, Frost MH, et al. The Mayo Clinic manuscript series relative to the discussion, dissemination, and operationalization of the Food and Drug Administration guidance on patient-reported outcomes. Value Health 2007;10(Suppl. 2):S59–63.
[9] Speight J, Barendse SM. FDA guidance on patient reported outcomes. BMJ 2010;340:c2921.
[10] Willke RJ, Burke LB, Erickson P. Measuring treatment impact: a review of patient-reported outcomes and other efficacy endpoints in approved product labels. Control Clin Trials 2004;25:535–52.
[11] Caron M, Emery MP, Marquis P, et al. Recent trends in the inclusion of patient-reported outcomes (PRO) data in approved drugs labelling by the FDA and EMA. PRO Newsl 2008;40:8–10.
[12] Goozner M. Patient reported outcomes poised for takeoff after final FDA guidance. The Pink Sheet Jan 25, 2010;72(004): Art No. 00720040011.
[13] Marquis P, Caron M, Emery M, et al. The role of health-related quality of life data in the drug approval process in the USA and Europe: a review of guidance documents and authorizations of medicinal products from 2006 to 2010. Poster presented at International Society for Pharmacoeconomics and Outcomes Research, Baltimore, MD, May 2011.
[14] Gondek K, Sagnier P, Gilchrist K, et al. Current status of patient-reported outcomes in industry-sponsored oncology clinical trials and product labels. J Clin Oncol 2007;25:5087–93.
[15] Mordin M, Lewis S, Gnanasakthy A, et al. Patient-reported outcomes in product development guidance. Value Health 2010;13:A17 (Abstract PMC21).
[16] Johnson JR, Williams G, Pazdur R. Endpoints and United States Food and Drug Administration approval of oncology drugs. J Clin Oncol 2003;21:1404–11.
[17] Mordin M, Clark M, Siersma C, et al. Impact of the FDA draft guidance on patient reported outcomes (PRO) label claims for approved drug products in the US: has it made a difference? Value Health 2009;12:A29 (Abstract PMC55).
[18] Kamal KM, Desselle SP, Rane P, et al. Content analysis of FDA warning letters to manufacturers of pharmaceuticals and therapeutic biologicals for promotional violations. Drug Inf J 2009;43:385–93.

EUROPEAN JOURNAL OF CANCER 45 (2009) 347–353

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejconline.com

Review

Patient-reported outcomes: Assessment and current perspectives of the guidelines of the Food and Drug Administration and the reflection paper of the European Medicines Agency

Andrew Bottomley[a,*], *Dave Jones*[b], *Lily Claassens*[a]

[a]*EORTC Quality of Life Department, EORTC Headquarters, Avenue E. Mounierlaan, 83/11, Brussels 1200, Belgium*
[b]*AstraZeneca, Alderly Park, Macclesfield, England, United Kingdom*

ARTICLE INFO

Article history:
Received 10 September 2008
Accepted 30 September 2008
Available online 14 November 2008

Keywords:
Patient-reported outcomes
Regulatory guidance
Health-related quality of life

ABSTRACT

Aims: Patient-reported outcomes (PROs) have recently gained greater credibility with regulatory bodies aiming to standardise their use and interpretation in RCTs, thereby supporting medicinal product submissions. For this reason, the United States (US) Food and Drug Administration (FDA) and the European Medicines Agency (EMEA) have released guidelines. This review paper provides an overview of the current perspectives and views on these guidelines.

Method: To evaluate the FDA and EMEA PRO guidelines, 47 expert responses to the FDA guidance were qualitatively reviewed. Two reviewers independently extracted data from these letters and checked these responses to warrant consistency and agreement in the evaluation process. A PubMed literature review was systematically examined to obtain supporting evidence or related articles for both the guidance documents.

Results: Generally, there is agreement between regulatory authorities and the research community on the contents of the FDA and EMEA PRO draft guidance. However, disagreements exist on significant philosophical topics (e.g. the FDA focuses more on conceptual models and symptoms than the EMEA) and design topics (e.g. the FDA is more restrictive on issues of recall bias, blinding of oncology trials and degrees of psychometric validation than researchers and the EMEA). This could influence the approval of PRO claims.

Conclusion: PRO guidance from the EMEA and FDA has been valuable, and has raised the profile and active debate of PROs in oncology. However, our review of the current opinion shows that there are controversial aspects of the guidance. Consequently, greater latitude should be given to how the guidance is interpreted and applied.

* *Corresponding author:* Tel.: +32 2774 1661; fax: +32 2779 4568.
E-mail address: andrew.bottomley@eortc.be (A. Bottomley).

0959-8049/$ - see front matter
doi:10.1016/j.ejca.2008.09.032

1. Introduction

In the recent years, the use of patient-reported outcomes (PROs) has increased significantly.[1] HRQOL measures involve subjective patient assessment or evaluation of important aspects of well-being[2] that are affected by current disease and/or treatment. Prominent examples of cancer-related HRQOL tools are the EORTC QLQ-C30 and the Functional Assessment of Cancer Therapy General (FACT G).[3,4]

Recently, the Food and Drug Administration (FDA) introduced the umbrella term PRO and attempted to standardise PRO use to provide a more systematic treatment-review process. For this reason, the FDA released draft guidance on PRO measures in February 2006: *Guidance for Industry. Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims*.[5] Also, the European Medicines Agency (EMEA) produced a reflective paper of regulatory guidance for the use of HRQOL measures in the evaluation of medicinal products in cancer in July, 2005.[6]

However, despite considerable effort to develop guidelines, it is not entirely clear what evidence the regulatory bodies would require in supporting claims when reporting PRO data. To date, established PRO measures are criticised by the regulatory agencies, leading to rejection of many PRO labelling claims.

The main objective of this paper is to review the current opinion relating to the guidance, and to make recommendations based on our review.

2. Materials and methods

The aim of the study is to provide an overview of the current opinion in relation to the two major guidance documents: the FDA Patient-Reported Outcomes draft (2006) and the EMEA HRQOL reflection paper (2005). A thematic qualitative approach was used to compare recommendations and requirements from each guidance. Written comments, invited by the FDA and submitted to the FDA web site (Table 1), have been reviewed. These reviewed responses related to the FDA guidance were also considered in relation to the EMEA document.

In order to seek evidence in support of the statements or recommendations within these regulatory guidance documents, a systematic literature search using PubMed was undertaken from January 1990 to December 2007. This searched for key words related to recurring issues which arose within the documents to identify scientific evidence to clarify debatable issues. All searches were restricted to English language articles only. In addition, literature references were checked to identify further evidence. Abstracts for major conferences, e.g. ASCO (2005–2008) and ESMO (2005–2008), were also reviewed.

3. Results

The FDA PRO guidance generated 47 written responses, totalling to 364 pages of comments, which are accessible on the FDA web site. These comments mainly came from professional groups in both academia and the pharmaceutical sector (Table 1). No documents were found on the EMEA web site about the EMEA HRQOL reflection paper, but the literature search identified several studies (discussed later) that looked at these guidelines. The search on PubMed and the search of grey literature generated additional documents referred to in the results below. The results are presented on the main thematic areas arising from the documents and from the responses to the documents. The comments' numbers refer to the authors who are listed in Table 1.

3.1. *Conceptual framework and end-point model*

According to the FDA guidance, PRO instruments must be based on an appropriate and clearly defined framework. This requires documentation of patient interviews, literature reviews and expert clinical opinion in order to support the concepts, domains and their associations (FDA, Section IV A, p. 7). The EMEA briefly noted the importance of incorporating the clinically relevant health-related domains of functioning that impact on HRQOL (p. 3).

Additionally, a submission to the FDA should be supported by an end-point model,[7] which displays an overview of the relevant end-points in an RCT and their relationships by mapping the treatment benefit and appropriate claims. Such a model should be hypothesis driven and incorporate a specific perspective,[8] i.e. the FDA prefers researchers to pre-specify expectations concerning the treatment impact, such as impact on disease symptoms, and the scales by which these outcomes will be measured. Some of the respondents who commented on this FDA opinion agree with a conceptual framework as a basis for PRO questionnaires (Comments 29, 33 and 36). Others suggested certain amendments, e.g. in its definition (Comments 6 and 38), or emphasised the distinction between HRQOL and symptoms (Comments 28 and 44). However, clearer guidance would be appreciated on the type of empirical data that need to support the conceptual framework (Comments 13 and 14), and two respondents recommended that the FDA should clarify that a conceptual framework precedes empirical analyses and may change during the validation process (Comments 15 and 20). Some respondents (Comments 10 and 34) questioned if a conceptual framework is needed in the manner the FDA stipulates (e.g. in a diagram).

3.2. *Patient involvement in instrument development*

The FDA aims to review instrument development to determine whether an adequate number of patients have supported the opinion that the specific items in the instrument are adequate and appropriate to measure the desired concept(s) (FDA, Section IV B, p. 10). No EMEA position is given on this issue. It is unclear what the appropriate number of patients involved in item generation ought to be according to the FDA. Eight respondents raised the question of how the FDA will evaluate if the sample size of patients, used in questionnaire generation, is adequate (Comments 8, 11, 15, 22, 24, 30, 39 and 43). One respondent (Comment 20) stated that item generation often involves small numbers of patients.

In addition, the characteristics of patients participating in early questionnaire development should match the charac-

teristics of the target population. The FDA plans to compare the patient population used in the PRO instrument development to the study populations enrolled in clinical trials (FDA, Section IV B, p. 9). The EMEA made no related statements on this matter. Ten respondents were of the opinion that population comparisons using the list of specifically age, sex, ethnic identity and cognitive ability are too specific (Comments 9, 19, 20, 21, 23, 29, 30, 34, 39 and 43). Many respondents argued that while the pertinent characteristics should be determined and compared, these might be different from those listed by the FDA (e.g. disease type and disease severity).

3.3. *Multi-domains versus single domains: making PRO claims*

The FDA and the EMEA stated that general PRO claims should be based on and supported by improvement in all domains (FDA, Section IV A, p. 8) or at least in the most important domains (EMEA, p. 3). Specific claims based on individual items may be proposed, though only if these single items are validated for this purpose and pre-specified in the Statistical Analysis Plan (SAP). Twelve respondents criticised or questioned this requirement in various ways. For example, patients may not be impaired in all domains (Comments 1, 43), and therefore, an improvement in the totality of domains is not deemed feasible. A flexible approach is preferred that requires improvement in the most important domains (Comment 15), and no trend towards worsening for the other domains (Comment 37). Furthermore, it was suggested that a single component could be the basis of a general claim (Comments 10, 29, 32 and 41), and one respondent stated it should be pre-specified in the Statistical Analysis Plan (Comment 34). Several respondents needed more details on this issue (Comments 13, 24, 34 and 39).

3.4. *Recall period*

An appropriate recall period is required when assessing the effects of treatment on oncology PROs. The FDA argues for the measurement of the current state (Section IV B, p. 11). In contrast, the EMEA made no statements on a recall period. There is a consensus among respondent views that averaging experience over a period of days does not necessarily invalidate measurement (Comments 12, 14, 15, 18, 19, 21, 23, 24, 32, 34, 39 and 40): e.g. it can be valuable to measure patient evaluation of change over time (Comments 1, 10 and 36). Besides, the current state assessments over a longer period of time may include bias as well: e.g. caused by response shift (Comment 11). Two respondents (Comments 20 and 32) agreed that a 'current state' approach gives rise to data being influenced to too high an extent by the extremes of patient state. Indeed the current state can also be influenced by the use of concomitant medications. Furthermore, symptoms are believed to be appropriate for the current state assessments while instrumental activities of daily living may not be appropriate for current state measurement (Comment 20). Several respondents considered that the guideline may be too prescriptive or too general (Comments 11, 12, 19, 21 and 26) or contradictory to the previous FDA advice (Comment 22). The choice of recall period should depend on the type of disease (Comments 12, 19, 20 and 43), the question asked (Comments 19, 24 and 43) and the situation related to patient health and disease (Comment 15). Moreover, the FDA opinion on a recall period was believed to be unsupported by consistent or sufficient evidence (Comments 12, 15, 23, 24, 41 and 43). However, three respondents did agree with the FDA proposal of the current state measurement (Comments 7, 17 and 28).

Previously published research addresses the accuracy of recall in pain patients[9–15] and shows variability in the results. Certain researchers claimed retrospective perception of change may not be accurate due to recall bias.[13,15] Other researchers considered recall of experience to be equivalent to assessment of the current state[14,9–11], concluding that measurement error and significant regression effect are the main concerns in momentary measurements.[11]

3.5. *Reliability and validity*

The FDA guidance emphasised test–retest reliability as the most important reliability type, while internal consistency may be used in the absence of the test–retest reliability (Section IV C, p. 18). EMEA did not address this topic. Eight respondents did not fully agree with the FDA: measuring test–retest reliability may be inappropriate or not feasible in some circumstances, and may be replaced by other reliability tests, e.g. internal consistency reliability (Comments 10, 11, 15, 23, 28, 29, 36 and 41). One respondent (Comment 6) noted that evidence is lacking for this FDA opinion.

Further, the FDA recommended that content-related validity should be addressed by providing evidence that items and response options are of a relevant and comprehensive nature with respect to the concepts that should be measured (Section IV C, p. 16). Evidence should contain a documentation of the issues abstracted from the literature; interview processes involving patients and healthcare providers (including interview transcripts) and information relating to the addition or deletion of items. Construct validity determines the extent to which items, domains and concepts relate to one another, supported by item-scale correlation analyses. Finally, predictive validity, the ability to predict the future outcomes through patient characteristics with prognostic value, is on the FDA list of psychometric properties (Section IV C, p. 16). However, it is questioned if this type of validity should be obligatory evidence in a submission to the agencies. Seven respondents agreed that flexibility is important with respect to the predictive validity (Comments 15, 18, 19, 21, 24, 39 and 43), since it may be unrealistic in some PROs, especially in new instruments. Our PubMed literature review showed the documentation of psychometric characteristics (e.g. construct validity) to be of great importance when trying to obtain FDA approval.[16] However, it would be appreciated if the FDA would emphasise that the demonstration of all measurement properties is an 'ideal' rather than a 'requirement' (Comment 34).

3.6. *Ability of PRO measures to detect change*

An instrument should detect changes in outcome measures if relevant clinical changes have occurred. For this reason,

evidence must be documented showing the degree to which the PRO instrument detected expected changes in values that are thought to have changed (part IV C, p. 18). This point was not debated significantly.

3.7. Interpretability of PRO measures

A Minimum Important Difference (MID) can be generated by applying a variety of methods, e.g. an anchor-based or a distribution-based approach or an empirical rule (FDA, Section IV C, p. 19). The EMEA stated that a MID 'should be based upon a combination of statistical reasoning and clinical judgment' (p. 5).

Respondents believe that MID should not be referred to as the only method for interpretability (Comments 34 and 35): some argued that it is not an exact science and has no clear evidence base (Comments 6 and 9). Three respondents explicitly agreed with the way the FDA proposes to establish a MID (Comments 14, 25 and 38). Some respondents advocated an anchor-based approach (Comments 8, 9, 14 and 41). Several respondents agreed on the additional value of input from patients or the clinical and research communities (Comments 25, 26, 29, 37 and 41), not using mathematical procedures alone (Comment 34). Although Sloan et al.[17] believe that combining these perspectives is favourable, they recognise that an even broader notion of HRQOL can be useful for the determination of clinically significant changes. Clarification is needed if MID refers to a 'between-group' change, a 'within-group' change or a 'within-patient' change (Comments 8, 12, 30, 35 and 36). The question arose as to when to use which approach: the MID or the responder analysis approach (Comment 15) and which one is the preferred methodology for the FDA (Comment 22). Joly et al.[18] note that the majority of advanced cancer RCTs (81%), published from 1994 to 2004, compute group differences. Nevertheless, they advocate defining a palliative response (i.e. observing the individual responder proportion), since HRQOL is the perception of an individual and not of a group. Finally, other respondents argued for the FDA to adopt a flexible approach towards the development of MID approaches (Comments 18, 21, 23 and 43), and some respondents noted that the MID can be influenced by many factors such as patient characteristics, the degree of the disease severity and finally how effective the therapy is (Comments 11, 14 and 39).

3.8. Blinding and randomisation related to PRO studies

According to the FDA (Section V A, p.23), open-label studies, in which patients and investigators are aware of the assigned therapy, are rarely credible. Open-label studies are also not recommended by the EMEA (p. 5). Many respondents challenged the FDA view: a majority believe that blinding in oncology clinical trials is hardly feasible in some circumstances (Comments 11, 12, 18, 22, 25, 32, 38, 39, 41 and 43) and required flexibility here (Comment 21). Current open-label studies are not by definition believed to provide invalid data (Comments 20, 23, 24, 26 and 43). It is even stated that non-blinded, naturalistic trials may give rise to more valid estimates than rigorously blinded or open trials.[19]

One of the five sources of bias as to why, up to date, HRQOL-based efficacy claims were disapproved by the FDA is believed to be the lack of randomisation.[20] In general, the procedure of randomisation seems to be commonly applied in the current oncology RCTs. Respondents and the EMEA released no significant statements.

3.9. Statistical concerns

A SAP should address the methodology of handling the missing data: a range of methods for doing this are listed by the FDA (Section VI D, p. 29). Also, the EMEA stressed the importance of discussing the missing data in the study design (p. 4). Many respondents replied in various ways, e.g. addressing inadequacy of worst-case scenario (Comments 18, 34, 37 and 43); complete case analysis (Comment 13); imputation methods (Comments 34, 36 and 41); the importance of advance methods such as mixture models or joint/shared parameter models (Comment 41) and the prevention of missing data by means of electronic collection by PRO methods (Comments 17, 28 and 44). Although a statistical correction can be applied, prevention through careful study design and execution is believed to be the best approach,[1] as well as ensuring, where possible, that the reasons for the missing data are captured. At least one respondent agreed on flexibility (Comment 18) in dealing with analysis. In addition, there was agreement about the pre-specification of methods in the SAP or protocol (Comments 23 and 38).

Further, the pre-specification of a sequence of testing, or a hierarchy of comparisons that need to be satisfied before others are considered for testing, is recommended in order to control for substantial increases in type I error (FDA, Section VI B, p. 27). According to the EMEA, multiplicity in PRO assessment may be overcome by the pre-specification of a subset of HRQOL domains, correction of *p*-values, hierarchical testing or global test procedures (p. 5). Three of five respondents advocated greater flexibility with respect to *post-hoc* or *ad-hoc* analyses. Such analyses are believed to offer additional value in identifying unanticipated patient benefit and in better clarifying results (Comments 15, 21 and 43). Other comments concerned the wish for clearer guidance (Comments 13 and 20).

The sample size used in a trial should depend on several factors, e.g. the number of end-points and the decision rule for success (FDA, Section VI B, p. 27). Generally, according to the EMEA (p. 5), the number of patients necessary to support the change in the primary end-point is sufficient to test for PRO change. However, Joly et al.[18] believe that most studies are not powered for PROs, since sample sizes are typically based on only one end-point. Three of five respondents noted that a sample size should be driven by primary PRO scales when PRO studies are conducted (Comments 32, 39 and 41).

3.10. Modification of instruments

The FDA announced that revised instruments should be viewed as different from the original, and therefore additional validation studies are recommended (Section IV D, p. 20). The EMEA guidance contains no statements on this topic. Many respondents argued that this FDA requirement is too restrictive (Comments 11, 14, 15, 23, 26, 35, 41 and 43) and may lead to the stagnation of instrument development (Comment 40).

Re-validation studies were found to be unnecessary in the case of relatively minor modifications, such as revisions in wording and differences in disease severity levels (Comments 1, 9, 12, 19, 24, 28, 32, 34, 39 and 44). Cognitive debriefing tests might be sufficient for these relatively small changes (Comments 11 and 34). It was advised that the level of modification should provide the best guide to the extent of re-validation that is required (Comments 13, 18, 20, 25 and 29).

3.11. *Translations*

The FDA guidance on PRO translations recommends instrument developers provide evidence that the methods and results of the translation process were adequate to warrant the validity of the responses. Accepted standards for translation and cultural adaptation must be applied to support their validity (FDA, Section IV D 5, p. 21). The EMEA released no translation statements. Respondents questioned what the generally accepted standards are (Comments 15 and 25), and expressed a wish for additional translation guidance (Comment 1). Five respondents doubted whether full validation should be required for each new translation (Comments 11, 22, 34, 36 and 40). Recently, Acquadro et al.[21] have stated that a multi-step approach gives rise to high-quality translation.

4. Discussion

In oncology phase III RCTs and registration trials, PROs are increasingly used for providing information about HRQOL in patients who undergo new treatments. Both the FDA and EMEA increasingly appear to be willing to accept PROs in support of medicinal labelling claims or in the evaluation of medical products such as cancer drugs.

The views of the FDA on PROs could be described as extensive, with detailed requirements and a restrictive nature. The EMEA has provided more global statements and broad advice which suggests that researchers should use the best available evidence and current standards in their trials. Another point of difference in PRO guidance from both agencies is the FDA's emphasis on the need for end-points that reflect direct consequences from disease and treatment and on the requirement for simple and easy to measure end-points such as symptoms, whereas the EMEA's focus goes beyond symptoms and includes HRQOL. Despite this, no major fundamental contradictions have appeared.

Although the attention on PROs has increased, PRO end-points have infrequently contributed to oncology product-approval to date. This was found in studies addressing regulatory approvals of oncology products with PRO (HRQOL) statements in their label claims.[22,23]

Also, former approvals with PRO claims do not reflect the current FDA and/or EMEA requirements or accepted current thinking. Consequently, no insight can be drawn from these earlier approvals, and therefore, they should not serve as examples for future submissions with PRO components.

Our review takes into account FDA and EMEA regulations as well as the perspectives of other key stakeholders, e.g. academia and pharmaceutical sectors. The guidance shows that it is important to provide sufficient documentation to support the PRO submission. Furthermore, systematic and ongoing correspondence with the regulatory authorities during the development process of a trial design and/or a PRO questionnaire is of major importance. Also, well-defined and hypothesis-driven end-points should be chosen that clearly reflect treatment benefit for the FDA. While the FDA appears to focus much on specific end-points such as symptoms, the EMEA appears more likely to accept domains such as overall health-related quality of life and functioning domains. Several issues remain unclear including the need for the FDA end-point model, describing the relationships among end-points, and a conceptual framework. Patient input in item generation should be demonstrated, but the extent to which interviews must be documented and the exact number of patients incorporated are unclear. The recall period chosen in the FDA guidance was heavily criticised by respondents. A substantial section of the research community argues that the FDA guidance is too prescriptive on this issue and is lacking evidence for much of the recommendations related to the recall period.

Regulatory agencies such as the FDA are likely to criticise the ability of the questionnaire to support certain label claims. Therefore, care should be taken with the psychometric evidence of each single scale by showing its validity and its ability to independently support a claim. Since improvement in all domains is frequently unrealistic, sponsors should propose specific claims, not merely broad claims, pre-specified in the protocol or the SAP. Several methods can be used in order to generate a MID, since the regulatory guidelines maintain flexibility in this matter. A preference exists to involve opinion from the clinical perspective, as supported by the EMEA and respondent statements. Therefore, an approach that integrates clinical and mathematical input should find support from the regulatory agencies. An acceptable statistical practice includes the pre-specification in the SAP of the way the missing data will be handled, the plan for coping with multiplicity and issues concerning sample size. Translation procedures required by the FDA PRO guidance are not explained in great detail, but following international standards should be adequate.

The FDA PRO regulatory guidelines are found to be stringent for well-established PRO or HRQOL measures with a long history of effective use and significant evidence of real world validity: it is questioned to what extent the existing questionnaires are supposed to meet the draft regulations, even with decades of supporting evidence (Comments 9, 15, 19, 23, 24, 26, 30, 34 and 37). An example of a widely used European questionnaire is the EORTC QLQ-LC13, a 13-item lung cancer-specific module – developed alongside the EORTC QLQ-C30[24] – capturing cancer-associated symptoms and therapy side-effects.[25] The generation of these measures included patient input, i.e. a certain number of patients for several development phases, and ensured population representation. Overall, they were developed to standards the FDA and EMEA note.[24,25] However, given the current guidelines, the FDA may incorrectly question the EORTC QLQ-C30 and the lung module and other major, well-established tools (e.g. FACT) given they were developed decades before the new guidelines were issued. Not all documentations the FDA may request will be maintained by all established instrument developers. In the

case of the EORTC QLQ-C30, there were no concerns relating to recall bias in thousands of trials involving 10,000s of patients carried out over several decades.[26] Hence, the FDA must take a much more liberal view of its guidance on such factors as present state assessment and recall period.

Our review has limitations. Specifically, qualitative evaluations of the data were made on the basis of submitted letters to the FDA web site. Some respondent comments may have included vague statements, and therefore are difficult to interpret or to include in this review. The use of two independent reviewers and a third arbitrator has limited the potential for bias, although an element of subjective interpretation was required. Furthermore, due to word limit restrictions, only the major themes were addressed.

In conclusion, broadly, oncology researchers and clinicians have welcomed the FDA PRO draft guidance and the EMEA HRQOL reflection paper. These documents are considered important steps towards the acceptation and appreciation of patient viewpoint, and the creation of significant evidence in the drug approval process. Nevertheless, continued dialogue and future FDA PRO guidance improvements on the key methodological issues raised in our review will help make PROs an important element in the fight to improve patients' HRQOL.

Conflict of interest statement

EORTC and Andrew Bottomley have received an unrestrictive academic research grant funding from AstraZeneca, UK. Dave Jones is an employee of AstraZeneca. Lily Claassens' Fellowship was in part funded by an academic grant from AstraZeneca.

Acknowledgements

We thank Dr. Bhash Parasuraman (AstraZeneca, USA) for her comments on our paper.

This research was funded by an educational academic grant from AstraZeneca, UK. In addition, part of the publication was supported by Grant Numbers 5U10 CA011488-37–5U10 CA011488-38 from the National Cancer Institute (Bethesda, Maryland, USA). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Cancer Institute.

Appendix A. Supplementary material

Supplementary data (Table 1) associated with this article can be found, in the online version, at doi:10.1016/j.ejca.2008.09.032, or at http://groups.eortc.be/qol/qolu_activities.htm.

REFERENCES

1. Bottomley A. Developing clinical trial protocols for quality of life assessment. *Appl Clin Trials* 2001;**10**:40–4.
2. Lipscomb J, Gotay CC, Snyder C. Introduction to outcomes assessment in cancer. In: Lipscomb J, Gotay CC, Snyder C, editors. *Outcomes assessment in cancer: measures, methods, and applications*. Cambridge, UK: Cambridge University Press; 2005. p. 1–14.
3. Bottomley A, Aaronson NK. European Organisation for Research and Treatment of Cancer International perspective on health-related quality-of-life research in cancer clinical trials: the European Organisation for Research and Treatment of Cancer experience. *J Clin Oncol* 2007;**25**(32):5082–6.
4. Lipscomb J, Gotay CC, Snyder CF. Patient-reported Outcomes in cancer: a review of recent research and policy initiatives. *CA Cancer J Clin* 2007;**57**:278–300.
5. Food and Drug Administration. Guidance for industry. Patient-reported outcome measures: use in medical product development to support labeling claims. US Food and Drug administration website, 2006 [cited 2007 April]. <http:www.fda.gov/cder/guidance/5460dft.pdf>.
6. European Medicines Agency. Committee for medicinal products for human use (CHMP). Reflection paper on the regulatory guidance for the use of health-related quality of life (HRQL) measures in the evaluation of medicinal products. European Medicines Agency website, 2005 [cited 2007 April]. <http://www.emea.europa.eu/pdfs/human/ewp/13939104en.pdf>.
7. Burke L, Rock EP, Powers JH. Patient-reported outcome instruments: overview and comments on the FDA draft guidance. US Food and Drug administration website, 2006 [cited 2007 Oct.]. <http://www.fda.gov/cder/present/DJA2006/Bruke_Rock.ppt>.
8. Lenderking W, Stewart M, Merikle E. Coping with the validation requirements of the FDA PRO guidance: the intersection of science and practice. In: *Conference proceedings, 14th annual conference of the international society for quality of life research, Toronto, Canada*; 2007. p. 8.
9. Bolton JE. Accuracy of recall of usual pain intensity in back pain patients. *J Pain* 1999;**83**:533–9.
10. Jamison RN, Raymond SA, Slawsby EA, McHugo GJ, Baird JC. Pain assessment in patients with low back pain: comparison of weekly recall and momentary electronic data. *J Pain* 2006;**7**(3):192–9.
11. Middel B, Goudriaan H, de Greef M, Stewart R, van Sonderen E, Bouma J, et al. Recall bias did not affect perceived magnitude of change in health-related functional status. *J Clin Epidemiol* 2006;**59**(5):503–11.
12. Erskine A, Morley S, Pearce S. Memory of pain: a review. *Pain* 1990;**41**(3):255–65.
13. Stone AA, Broderick JE, Shiffman SS, Schwartz JE. Understanding recall of weekly pain from a momentary assessment perspective: absolute agreement, between- and within-person consistency, and judged change in weekly pain. *Pain* 2004;**107**:61–9.
14. Brauer C, Thomsen JF, Loft IP, Mikkelsen S. Can we rely on retrospective pain assessments? *Am J Epidemiol* 2003;**157**:552–7.
15. Gendrau M, Hufford MR, Stone AA. Measuring clinical pain in chronic widespread pain: selected methodological issues. *Best Pract Res Clin Rheumatol* 2003;**17**(4):575–92.
16. Revicki DA, Gnanasakthy A, Weinfurt K. Documenting the rationale and psychometric characteristics of patient reported outcomes for labeling and promotional claims: the PRO Evidence Dossier. *Qual Life Res* 2007;**16**:717–23.
17. Sloan JA, Frost MH, Berzon R, et al. The clinical significance of quality of life assessments in oncology: a summary for clinicians. *Support Care Cancer* 2006;**14**:988–98.
18. Joly F, Vardy J, Pintilie M, Tannock IF. Quality of life and/or symptom control in randomized clinical trials for patients with advanced cancer. *Ann Oncol* 2007;**18**(12):1935–42 [Epub 2007].

19. Marquis P, Arnould B, Acquadro C, Roberts WM. Patient-reported outcomes and health-related quality of life in effectiveness studies: pros and cons. *Drug Develop Res* 2006;**67**:193–201.
20. Rock EP, Scott JA, Kennedy DL, Sridhara R, Pazdur R, Burke LB. Challenges to use of health-related quality of life for Food and Drug Administration approval of anticancer products. *J Natl Cancer Inst Monogr* 2007(37):27–30.
21. Acquadro C, Conway K, Hareendran A, Aaronson N. European Regulatory Issues and Quality of Life Assessment (ERIQA) Group Literature review of methods to translate health-related quality of life questionnaires for use in multinational clinical trials. *Value Health* 2008;**11**(3):509–21. Epub 2007.
22. Rock EP, Kennedy DL, Furness MH, Pierce WF, Pazdur R, Burke LB. Patient-reported outcomes supporting anticancer product approvals. *J Clin Oncol* 2007;**25**(32):5094–9.
23. Coombs JH, McBurney CR, Gondek K, Copley-Merriman K. Evidence of patient reported outcomes in labelling for oncology products: evaluation of United States and European labels. *Qual Life Newslett* 2002:29.
24. Aaronson NK, Ahmedzai S, Bergman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. *J Natl Cancer Inst* 1993;**85**(5):365–76.
25. Bergman B, Aaronson NK, Ahmedzai S, Kaasa S, Sullivan M. The EORTC QLQ-LC13: a modular supplement to the EORTC core quality of life questionnaire (QLQ-C30) for use in lung cancer clinical trials. EORTC study group on quality of life. *Eur J Cancer* 1994;**30A**(5):635–42.
26. Bottomley A. The journey of health-related quality of life assessment. *Lancet Oncol* 2008;**9**(9):906.

VALUE IN HEALTH 15 (2012) 443–448

Available online at www.sciencedirect.com

SciVerse ScienceDirect

journal homepage: www.elsevier.com/locate/jval

Reasons for Rejection of Patient-Reported Outcome Label Claims: A Compilation Based on a Review of Patient-Reported Outcome Use among New Molecular Entities and Biologic License Applications, 2006–2010

Carla DeMuro, MS[1,], Marci Clark, PharmD[1], Margaret Mordin, MS[1], Sheri Fehnel, PhD[1], Catherine Copley-Merriman, MS, MBA[1], Ari Gnanasakthy, MSc[2]*

[1]*RTI Health Solutions, Research Triangle Park, NC, USA;* [2]*Novartis Pharmaceuticals Corporation, East Hanover, NJ, USA*

A B S T R A C T

Objectives: Previous analyses of patient-reported outcome (PRO) label claims concentrated only on successful label claims. The goal of this research was to explore the reasons why PRO label claims were denied and to compile regulatory feedback regarding the use of PROs in clinical trials. **Methods:** By using the Food and Drug Administration's Drug Approval Report Web page, all new molecular entities and biologic license applications approved between January 2006 and December 2010 were identified. For identified drug products, medical review sections from publicly available drug approval packages were reviewed to identify PRO end-point status and any Study Endpoints and Label Development team comments. **Results:** Of the 116 new molecular entities and biologic license applications with accompanying drug approval packages identified and reviewed, 44.8% of the products included PROs as part of the pivotal studies; however, only 24.1% received PRO label claims. Primary reasons for denial included issues of fit for purpose, issues of study design, data quality or interpretation, statistical issues, administrative issues, and lack of demonstrated treatment benefit. **Conclusions:** Based on drug approval packages, nearly half (45%) of new molecular entitity/biologic license application products in the years 2006 to 2010 included PROs in the clinical trials supporting their approval, yet this rate is not reflected by claims granted. Understanding the nature of PRO claims granted under the current regulatory guidance is important. In addition, a clear understanding of denied claims yields valuable insight into where sponsors may improve implementation of PROs in clinical trials and submission of PRO evidence to increase the likelihood of obtaining PRO label claims.

Keywords: label claims, patient-reported outcomes, rejection.

Introduction

Patient-reported outcomes (PROs) allow the voice of the patient to emerge within the context of a clinical trial or observational study and provide valuable insight into the patient experience beyond that which can be measured by clinical indices alone. In some diseases or conditions of interest, a PRO may be the sole source of data from which drug efficacy can be measured, whereas in others it may provide supplementary information on how the disease and its treatment impact patients' functioning and well-being.

PRO use is particularly common for products developed to treat chronic, disabling conditions where the intention is not necessarily to cure but to ameliorate symptoms, facilitate functioning, or improve quality of life. PROs are the primary end points in clinical trials evaluating drug products for disease areas such as irritable bowel syndrome, migraine, and pain. PROs provide key supportive data in many other disease areas, such as insomnia, asthma, and psychiatric disorders. In oncology, PROs are commonly used to assess both treatment benefits and toxicity to fully evaluate the impact of treatment on health-related quality of life (HRQOL). PROs can also be used in clinical trials to assess treatment satisfaction, compliance, and caregiver burden [1].

Sponsors (i.e., pharmaceutical or biotechnology companies developing a new product) may choose to include a PRO end point to support a label claim, to provide data supporting the primary endpoint, or as a source of data for communication and market-access strategies. Regardless of the reason for a PRO's inclusion in a clinical trial, it is unique in that it captures the viewpoint of the patient without input from others.

Willke and colleagues [2] conducted a review of drug labels to understand the use of PROs compared with other trial end points. That research identified the inclusion of PROs as efficacy end points in approximately 30% of all labels reviewed between 1997 and 2002. In 2006, the Food and Drug Administration (FDA) released a draft guidance for use of PROs in clinical trials, followed by a final guidance in 2009, *Guidance for Industry: Patient Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims* [3], providing a blueprint for the use of PROs in clinical trials. The guidance documents were intended to influence the appropriate development, validation, and use of PRO measures to facilitate a positive regulatory review in support of label claims.

* *Address correspondence to:* Carla DeMuro, RTI Health Solutions, 200 Park Offices Drive, Research Triangle Park, NC 27709, USA.
E-mail: demuromercon@rti.org.

1098-3015/$36.00 – see front matter

http://dx.doi.org/10.1016/j.jval.2012.01.010

The Study Endpoints and Labeling Development (SEALD) team coauthored the PRO guidance in collaboration with other colleagues from the FDA Center for Drug Evaluation and Research, the Center for Biologics Evaluation and Research, and the Center for Devices and Radiological Health. SEALD acts as an advisory board to the 17 Office of New Drugs reviewing divisions within the FDA and provides guidance pertaining to the development and validation of study end points, clinical study protocol design, analysis, and interpretation of study end points to support drug development, labeling, and promotion.

According to the guidance, a claim is defined as a statement of treatment benefit. Furthermore, a claim can appear in any section of a medical product's FDA-approved labeling or in advertising and promotional labeling of prescription drugs and devices.

Since its release to the public, much interest has been paid to the impact of this guidance document on the use of PROs and the acceptance of PRO-based label claims [1,4,5,6]. Gnanasakathy and colleagues [1] built on the work previously conducted by Willke and colleagues [2] and reported the frequency of PROs in recently approved drug labels. Specifically, these authors found that PRO claims were granted for approximately 24% of all labels reviewed between January 2006 and December 2010.

To date, however, no formal review has been undertaken to examine PRO measures included in drug approval packages (DAPs) but not appearing in approved labeling. Hence, there is no compilation of feedback on the use of these PROs either by industry or by regulatory authorities. Examination of these submissions may provide an insight into the appropriate utilization of PROs by sponsors in clinical studies and additional guidance for preparing evidence dossiers. This information may also provide regulators with an overview to assess consistency in response across reviewing divisions. Therefore, the purpose of this research was to review the criticisms targeted at PRO end points for all new molecular entities (NMEs) and biological license applications (BLAs) from 2006 through 2010 that utilized PROs in their clinical trials supporting their approval but did not receive labeling claims for these measures.

Methods

Data collection methods for this research are fully described elsewhere [1]. Briefly, the FDA Drug Approval Reports Web page was used to review new drugs that were approved in the United States from January 2006 through December 2010, including only those products classified by the Center for Drug Evaluation and Research as NMEs or BLAs. Any product containing substances previously marketed with a different brand name or set of indications, as a different dosage form or strength, or as a combination product of previously marketed entities was excluded.

Once products were identified, DAPs and approved product labels were reviewed, and information was retrieved from the medical review, summary review, cross-discipline team leader review, and other review sections from the DAP as well as from the Indication and Clinical Studies section of the approved product label. The DAPs were located on the FDA Web site Drugs@FDA (www.accessdata.fda.gov). The following information was collected, as publicly available, for each US drug product identified:

- Brand name
- Generic name
- Date of approval
- Applicant
- Label indication
- Utilization of PROs
- PROs mentioned in the DAP but not appearing in the label
- Evidence of claims sought but not granted
- Significance of the PRO results
- Division reviewer or SEALD reviewer feedback

Statistical analysis consisted of frequencies and cross-tabulations of measured characteristics. Calculations were performed by using Microsoft Excel 2007. For analysis purposes, if a PRO appeared in the DAP, it was considered to be an attempt to seek a PRO label claim, despite sponsor intent, unless specifically noted otherwise.

Results

A total of 156 new drugs were approved between January 2006 and December 2010. Of these, 33 were generic products and were excluded from our analysis, as were 4 new products that were approved but had no data available on the FDA Web site at the time of review and three others were registered under multiple names so were considered single entities. Therefore, this review includes 116 products.

Of the 116 products reviewed, 52 (44.8%) included PROs as part of the pivotal studies; however, only 28 of the 116 (24.1%) received at least one PRO claim [1]. A total of 26 products were identified as having been denied a PRO label claim. For the purposes of analysis, this included any product that had a PRO included in the DAP, regardless of the sponsor's intention, because it was not always possible to determine whether a claim had been sought or whether PRO data had been collected for other reasons. A subset of products (n = 6) received some or partial PRO labeling while other requested PRO claims were denied within the same submission. These six products were Azilect, Chantix, Letairis, Ampyra, Bepreve, and Egrifta. Table 1 provides a listing of all 26 products described in this review, arranged by the FDA division that granted drug approval.

The filings for these 26 products included a wide range of PRO measures, for example, symptom diaries, event logs, measures of functioning and disability, symptom assessments (e.g., fatigue and pain), disease-specific measures of HRQOL, generic assessments of HRQOL, and utility measures. Table 2 provides an alphabetical listing of measures specified in the DAPs but not appearing in the approved labeling.

To determine the rationale behind decisions to reject PRO claims from the label, data specific to PROs mentioned in the DAP

Table 1 – Products with at least one claim denied by FDA reviewing division.

FDA reviewing division	Products reviewed
Anesthesia, analgesia, and rheumatology products	Chantix, Ilaris
Antiinfective and ophthalmology products	Lucentis, Bepreve
Biologic oncology:	Vectibix
Cardiovascular and renal products	Letairis, Samsca
Dermatology and dental products	Stelara
Drug oncology	Dacogen, Zolinza, Torisel, Ixempra kit, Treanda, Istodax, Jevtana
Gastroenterology products	Vpriv, Elaprase, Relistor
Medical imaging and hematology products	Promacta
Metabolism and endocrinology	Januvia, Egrifta, Somatuline
Neurology products	Azilect, Ampyra
Psychiatry	Invega, Pristiq

FDA, Food and Drug Administration.

Table 2 – Alphabetical listing of measures with claims denied.

Measure
Body Image Impact Module
Borg's Dyspnea Index
Caregiver Outcomes Assessment
Child Health Questionnaire–Child Form
Child Health Questionnaire, Parent Completed 50-Item Scale
Childhood Health Assessment Questionnaire
Chronic Idiopathic Thrombocytopenic Purpura Symptoms
Constipation Distress
Dermatology Life Quality Index
European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire–Chronic Lymphocytic Leukemia 25
European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire–C30
EQ-5D
Functional Assessment of Chronic Illness Therapy Fatigue Scale
Functional Assessment of Cancer Therapy–Breast Symptom Index
Health Assessment Questionnaire–Disability Index
Hospital Anxiety and Depression Scale
Hunter Syndrome–Functional Outcomes for Clinical Understanding Scale
Hyponatremia Disease Specific Survey
Itch VAS
McGill-Melzack Present Pain Intensity scale
Modified Cigarette Evaluation Questionnaire
Multiple Sclerosis Walking Scale–12
Opioid Withdrawal Symptoms (modified Himmelsbach)
Pain Numerical Rating Scale (0–10 scale)
Parkinson's Disease Quality of Life Scale
Patient Impression of Change in Bowel Status
Patient Reports of Bowel Consistency and Difficulty
Pruritis relief VAS
Quality of Life assessments by proxy
Short form-36
SF-36 Physical Functioning Scale
Sleep VAS
Subject Global Impression of Change
Symptoms and Quality of Life in Schizophrenia
Visual Function Questionnaire–25
Work Limitations Questionnaire

EQ-5D, EuroQol five-dimensional questionnaire; SF-36, Short Form 36 Health Survey; VAS, visual analogue scale.

(but not appearing in the labeling) were extracted for further examination. The following coding convention was created and applied by a single rater to categorize the FDA reviewer's (division or SEALD) noted concerns regarding the PRO measure:

1. Fit for Purpose: lack of evidence of content validity (e.g., lack of link between concept and claim, insufficient documentation of validation in population of interest, and full constellation of symptoms not measured), recall period, or lack of evidence of proper translation or cross-cultural validation;
2. Study Design, Data Quality, or Interpretation of Results: issues of potential bias (open-label design, etc.), clinical meaningfulness, missing data, attrition rates, or improper completion;
3. Statistical Analysis: no adjustment for multiplicity or inappropriate or missing statistical analysis plan;
4. Administration Considerations: lack of documentation for training or instruction in use of measure or copy of measure not provided to the FDA; and
5. No Treatment Benefit: not supportive of treatment benefit, improvement in certain symptoms but worsening in others, lack of statistical significance, or FDA disagreed with sponsor.

Examination of the DAP for each product provided differing levels of detail regarding why a measure was not included in the approved labeling. Reasons for this included the proprietary nature of labeling discussions between sponsor and agency as well as differences between products receiving a review by SEALD. Detailed feedback for each submission, by product, is provided in the appendix in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2012.01.010.

"Fit for purpose" issues were the primary reasons for the denial of PRO labeling claims, accounting for more than 38% of regulatory feedback. A PRO measure that has been recognized by the FDA as appropriate to support claims in a specific context (i.e., the measure meets the qualifications for supporting claims outlined in the PRO guidance, specific to a study population and protocol and to the claim sought/hypothesis tested) is described as "fit for purpose" by the FDA.

As cited in the DAPs for 14 individual products, the FDA specifically questioned the content validity and/or validity of instruments in general, rationale in support of recall periods, and evidence of appropriateness of translations for use in multinational studies. This feedback was consistent, especially in regard to validity. Of the 14 products that fell within this category, 8 were noted to have potential issues with the validity of the PRO measure for the intended purpose. A SEALD review of the use of the SF-12 Health Survey (SF-12) and the Hyponatremia Disease Specific Survey as secondary end points in pivotal studies of Samsca provides an illustrative example. In these studies, the sponsor included the SF-12 and justified the use of the tool by pointing out that hypo-

natremia presents in a broad range of disease areas and that both the physical and mental component scores of the SF-12 were used. Reviewer feedback, however, noted: "The SF-12 was developed as a generic health status instrument for the general population and not as a symptom assessment tool or HRQoL tool in patients with hyponatremia. As such the instrument is not effective as an assessment of treatment benefit." Regarding the Hyponatremia Disease Specific Survey, the SEALD reviewer explained that "the information submitted by the sponsor concerning the psychometric properties of the HDS do not address the content validity and therefore do not support the use of the instrument." Similar feedback was provided in the Torisel review where FDA reviewers noted that the "applicant did not provide evidence of validation of the EQ-5D [EuroQol five-dimensional questionnaire] in the RCC [renal cell carcinoma] population. It was used in a setting for which it was not designed, and more frequently than intended."

Issues of study design, data quality, or interpretation of results was the second largest identified category and accounted for approximately 27% of the feedback for denied claims. In this category, reviewers questioned the clinical meaningfulness of patient responses, noted issues of bias introduced by open-label study designs, and commented on missing data/dropout rates and other indicators of data quality. These concerns were identified for nine individual products. Regulatory feedback on Zolinza and Torisel was illustrative of these points. The reviewer for the Zolinza submission stated, "PROs cannot be reliably measured in open label studies . . . a 3-point improvement was considered clinically significant, but the review does not state whether the proportion of patients obtaining this level of relief was clinically meaningful." Missing data and potential for bias were noted in the Torisel review. Neither Zolinza nor Torisel was granted PRO-related claims.

Statistical considerations that generated regulatory criticisms included lack of or inappropriate statistical analysis plans such as no planned adjustments for multiplicity. This issue is clearly described in the regulatory review of Azilect. The reviewer noted, "I cannot draw serious conclusions about the efficacy of these [PRO] end points because of issues of multiplicity whereby the sponsor did not make statistically appropriate adjustments for these multiple comparisons . . ." despite significant findings on the Parkinson's Disease Quality of Life scale in favor of Azilect. Although it is unknown how this adjustment may have impacted the result and subsequently the label claim, the expectations of the reviewing division are well documented.

In addition, administrative considerations impacted agency reviewer decision making. Concerns were noted regarding the lack of appropriate documentation describing training procedures, administration of the tool, and inadequate descriptions of measures. Examples of such concerns included the SEALD reviews of Egrifta where reviewers noted a missing user's manual, lack of description of the Caregiver's Outcome Assessment for Torisel, and confusion regarding patient instructions for using an itch visual analogue scale for Stelara.

A final category grouped agency reviewer feedback on PRO measures where discrepancies occurred between the agency and the sponsor regarding whether a measure appropriately demonstrated treatment benefit. Feedback in this category ranged from a straight-forward assessment of no demonstrated statistical difference between active treatment and placebo (e.g., Letairis and Relistor) to more detailed discussions of failure to demonstrate treatment benefit when some symptoms improved while others showed worsening (e.g., Chantix).

Figure 1 depicts the percentage of claims denied by each analytic category of reasons for rejections, and Table 3 describes regulatory feedback by product.

Case studies of each drug submission are detailed in the Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2012.01.010. Differing levels of information were provided in each DAP; likewise, review formats were somewhat inconsistent. Therefore, the level of detail extracted from the submissions varies by product. For the purposes of this review, it was assumed that the sponsor sought a claim based on the PRO(s) referenced in the DAP unless otherwise specified.

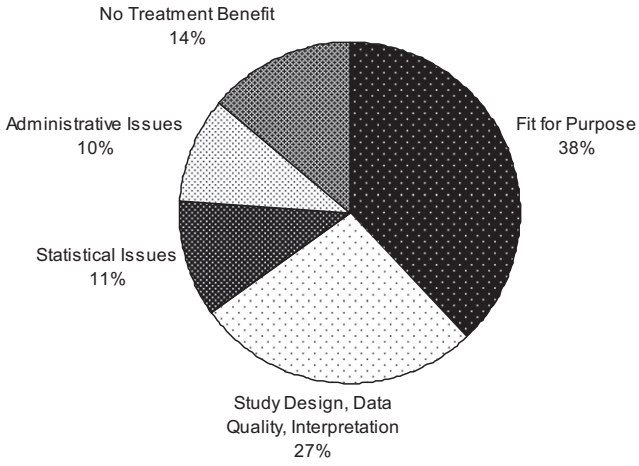


Fig. 1 – PRO label claim denials.

Discussion

To our knowledge, this is the first comprehensive compilation of FDA feedback on the use of PROs in clinical trials in support of label claims since the release of the draft FDA guidance in 2006. Reasons for rejection of claims varied, but the majority focused on whether a measure was fit for the purpose for which it was used and issues of study design, data quality, or interpretation of PRO results. Most denials and critical discussions were consistent with the spirit of the PRO guidance. The final PRO guidance places strong emphasis on interpreting PRO data and on developing PRO measures. Instrument validity, in particular content validity, is discussed in detail in the PRO guidance. The guidance notes that other measurement properties will not be considered until evidence of this property has been appropriately determined. Reviewers emphasized this in their criticisms in a number of product reviews, including several that utilized generic measures.

Concerns with study design and interpretation of PRO data persist. For example, reference to minimal important difference was removed from the final guidance and replaced with a discussion of individual responses to treatment or responder definitions. This change, however, does not completely address the issue of demonstrating a clinically relevant change. Clinical trial considerations are addressed in the guidance, but these issues do not always have a solution that is practical for all clinical trial conditions (e.g., single-arm study design in oncology studies).

Statistical considerations also remain paramount to obtaining PRO claims. Responses to submissions clearly demonstrate that PROs must be treated with the same rigor as other clinical end points. Prospective, adequate statistical analysis plans must be developed to address issues such as multiplicity and methods for dealing with missing data.

Importantly, as this review period is inclusive of the release of both the draft and final guidance documents, the level and type of documented feedback provided to the public by the FDA is inconsistent. First, the level of review varied across submissions. Not all submissions received a review from SEALD, because this group acts on a consultancy basis. Submissions with a SEALD review (e.g., Stelara, Chantix, Samsca, and Egrifta) received very detailed

Table 3 – Category of denial by product.

Product	Fit purpose	Study design, data quality, interpretation	Statistical issues	Administrative issues	No treatment benefit
Azilect			X		
Chantix	X				X
Dacogen*					
Luncentis	X				
Elaprase					X
Vectibix		XX	XX		
Zolinza		XX			
Januvia					X
Torisel	X	XXXX	X	X	
Letairis	X		X		X
Somatuline	X				X
Ixempra		XX			X
Relistor	XX				X
Samsca	XX				
Ilaris		X	X		
Stelara	XXXXXX			X	
Bepreve	X			X	X
Isodax	XX	XX		XXX	
Ampyra	X	X			X
Jevtana	XXX	XXXX		X	X
Egrifta	XXXX				
Invega			X		
Pristiq	X		X		
Treanda*					
Promacta*					
Vpriv		X			

* No information provided in the drug approval package. X, analytical category for denied claim.

comments and recommendations. Details on other submissions were much more difficult to discern and were found embedded within the medical review or cross-team leader review. Comments from the SEALD review of Egrifta illustrated the difficulties facing both industry and regulatory bodies in the review of studies utilizing PROs that were planned and executed prior to the release of the draft guidance. Specifically, the SEALD reviewer expressed reservations with respect to the content validity of the Body Image Impact Module, which does not meet the new standards articulated in the guidance; the reviewer stated that the instrument should not be recommended by FDA for future drug development, yet a claim was still granted. It is worth noting that the PROs evaluated in the Egrifta clinical trials had been incorporated with prior input from the FDA in advance of the final guidance. As experience with the guidance matures and both industry and regulatory bodies acclimate, such conflicts are expected to become less frequent.

Other inconsistencies in regulatory responses may be attributable to differences between reviewing divisions. For example, in some situations, a generic measure (e.g., the Short Form 36 Health Survey's physical component score) was accepted as a suitable end point by one reviewing division and rejected by another because of a lack of specificity. In addition, differences in the perceived acceptability of PRO measures to support labels claims may exist across FDA divisions. At the time of this analysis, no PRO claims have been granted by the oncology division, but PRO data for some oncology drugs (e.g., Dacogen) appear to demonstrate significant results regarding impact on HRQOL or symptoms such as fatigue and dyspnea. Although it is likely that these data were in some way confounded by the nature of the trial, sponsors would better understand the position of this division if details were provided in the reviews. Disclosing the measures used by sponsors and the reason for criticisms, if any, would greatly assist sponsors in refining their internal decision-making processes to include the right instrument to measure the right concept.

Several limitations should be noted for this review. First, for practical reasons we limited review of products to those classified as NMEs and/or BLAs. As such, products seeking approval for new indications were not included in our review. There may be instances where these submissions also have rejected PRO claims. A limitation of this analysis is that it is not clear, because of the confidential nature of labeling discussions, whether the comments by the FDA were for claims actively requested by sponsors or whether they were comments made in some other regard. PRO instruments are included in drug submissions not only for label claims but also to provide supportive data to the primary end point, to provide data requested by the FDA or the European Medicines Agency [5], for publication purposes, or to satisfy market-access needs (utility assessments). Unless actively seeking a label claim, the sponsor is unlikely to invest in new instruments to meet the standards outlined in the FDA PRO guidance. Therefore, although this analysis provides sponsors a means with which to assess and support the quality of their PRO strategies, our analysis is unlikely to be a measure of the quality of submissions targeted at PRO label claims to the FDA, because often the lack of access to a detailed response from the agency made it difficult to discern the rationale for these types of decisions.

Conclusions

The use of PROs as clinical trial end points continues to be widespread, with more than 45% of all NME or BLA submissions between 2006 and 2010 utilizing these instruments in some capacity [1]. Despite the commonality of PRO inclusion, rejection rates for PRO claims remain high. PRO label claims are denied for various reasons, some of which are addressed by the FDA in its PRO guidance. Although the learnings from this research are limited by the amount of information publicly available, review of denied claims

may provide an insight into how sponsors could improve the implementation of PROs in clinical trials and the level of PRO evidence submitted to increase the likelihood of obtaining PRO label claims. Such continuous learning and combined efforts between sponsors and regulatory bodies will allow the patient's voice to be heard in the drug development process.

Acknowledgments

We gratefully acknowledge the research assistance of Emily Evans in the development of this manuscript. We also gratefully acknowledge Lynda Doward and Jennifer Petrillo for review of the manuscript.

Source of financial support: This study was supported by Novartis Pharmaceuticals.

Supplemental Materials

Supplemental material accompanying this article can be found in the online version as a hyperlink at http://dx.doi.org/10.1016/j.jval.2012.01.010 or, if a hard copy of article, at www.valueinhealthjournal.com/issues (select volume, issue, and article).

REFERENCES

[1] Gnanasakathy A, Mordin M, DeMuro C, et al. A review of patient-reported outcomes labels in the US: 2006–2010. Value Health 2012;15:437–42.
[2] Willke RJ, Burke LB, Erickson P. Measuring treatment impact: a review of patient-reported outcomes and other efficacy end points in approved product labels. Control Clin Trials 2004;25:535–52.
[3] US Department of Health and Human Services. Guidance for industry: Patient-reported outcome measures: use in medical product development to support labeling claims. December 2009. Available from: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM193282.pdf. [Accessed January 14, 2011].
[4] Caron M, Emery MP. PRO labeling claims in antineoplastic agents. Value Health 2010;13:pA45,PCN111.
[5] Mordin MM, Clark M, Siersma CA, et al. Impact of the FDA draft guidance on patient reported outcomes (PRO) label claims for approved drug products in the US: has it made a difference? Presented at: the ISPOR 14th Annual International Meeting, May 2009, Orlando, Florida, USA. Available from: http://www.ispor.org/RESEARCH_STUDY_DIGEST/details.asp. [Accessed September 21, 2011].
[6] Viswanathan S, Gemmen, EK, Bharmal M. Evaluating central nervous system drug labels for patient reported outcomes. Presented at: the ISPOR 14th Annual International Meeting, May 2009, Orlando, Florida, USA. Available from: http://www.ispor.org/RESEARCH_STUDY_DIGEST/details.asp. [Accessed September 21, 2011].

Volume 12 • Number 8 • 2009
VALUE IN HEALTH

A Concept Taxonomy and an Instrument Hierarchy: Tools for Establishing and Evaluating the Conceptual Framework of a Patient-Reported Outcome (PRO) Instrument as Applied to Product Labeling Claims

Pennifer Erickson, PhD,[1] Richard Willke, PhD,[2] Laurie Burke, RPh, MPH[3]

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

ABSTRACT

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

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

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

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

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

Introduction

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

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

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

The ability to identify and codify this relationship has several advantages to sponsors, regulators, as well as to outcomes researchers more broadly. First, the sponsor and FDA need to understand the complexity of the concept in the desired claim because it will determine the adequacy of the instrument used to

Address correspondence to: Pennifer Erickson, Department of Public Health Sciences, Penn State College of Medicine, 1316 Deerfield Dr., State College, PA 16803, USA. E-mail: pae6@psu.edu

10.1111/j.1524-4733.2009.00609.x

  1098-3015/09/1158 1158–1167

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

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

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

Methods for Developing the PRO Concept Taxonomy and PRO Instrument Hierarchy

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

The second step was to validate the taxonomy and hierarchy by evaluating the labeling for the 142 new products approved by FDA from January 2003 through December 2007; labeling for 36 products reported at least one PRO. The PRO concepts and instruments found in labeling for the 1997–2007 period can be found at: http://www.ispor.org/Publications/value/ViHsupplementary/ViH12i8_Erickson.asp. The same methods were used for this review as for that of the 1997–2002 labeling. Third, we broadened the scope of our evaluation of PRO instruments to include formal scales beyond those that appeared in the new product labeling using information from the On-Line Guide to Quality-of-Life Assessment (OLGA) [6,22]. OLGA's comprehensive database includes information on thousands of instruments that are of potential relevance for supporting a claim of treatment benefit. Based on selection criteria designed to identify instruments of diverse conceptual content and measurement structures, the conceptual frameworks of 25 instruments were formally evaluated. This step provided assurance that the taxonomy and hierarchy would be relevant not only to instruments used in previous labeling, but also to those that might appear after 2007.

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

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

The PRO Concept Taxonomy

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

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

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

Each family, whether generic or specific, comprises at least one singular concept that both patients and their health-care decision-makers could consider to be a meaningful goal of treatment benefit, for example, pain intensity. Singular concepts may have low-level singular concepts that are considered to be too basic for use in labeling, for example, ability to cut meat. A

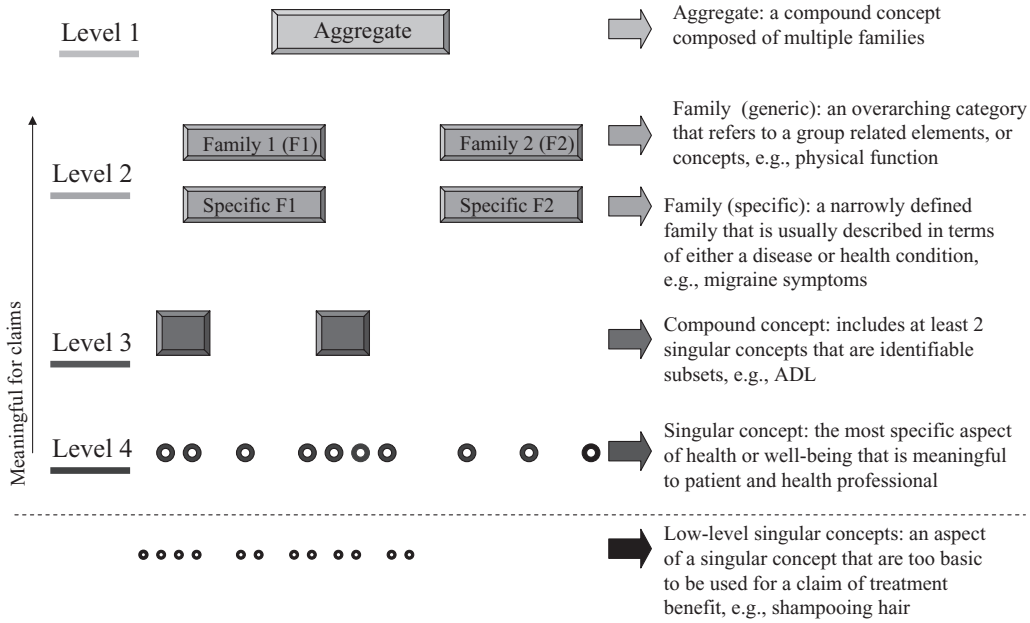


Figure 1 Patient-reported outcome concept taxonomy: depicts relationships between concepts. ADL, activity of daily living.

compound concept is defined as consisting of at least two singular concepts; for example, the concept "basic activities of daily living" typically includes bathing, toileting, transferring, and dressing.

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

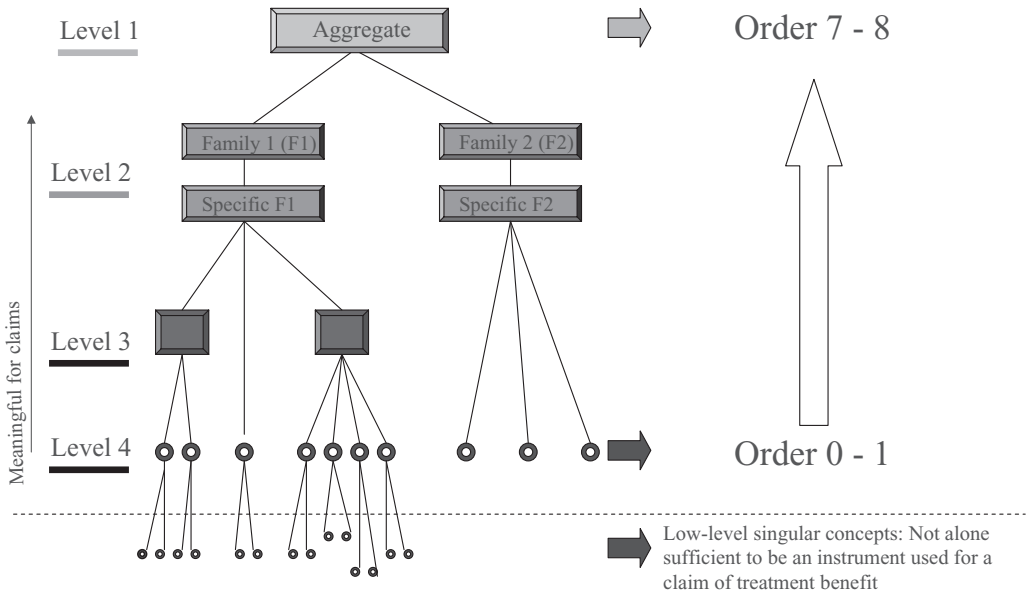


Figure 2 Patient-reported outcome instrument hierarchy: depicts concepts, measurement structure, and relationship to hierarchy.

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

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

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

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

The PRO Instrument Hierarchy

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

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

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

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

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

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

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

Table 1 PRO instrument hierarchy for classifying PRO instruments according to their taxonomic and measurement structures, with examples of PRO instruments and statements of treatment benefit from existing prescription drug labeling

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

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

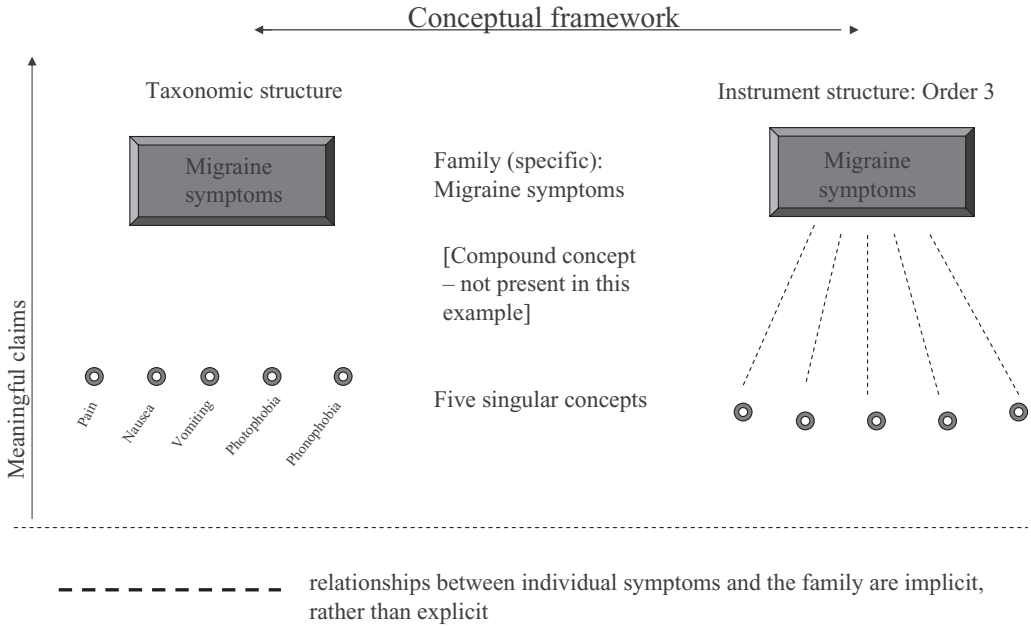


Figure 3 Intended statement of treatment benefit: reduce symptoms of migraine.

more than one family; these instruments can be used to measure both the concepts and the families. In addition, Order 7 instruments may be used to measure the concept represented by the aggregate score.

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

Depicting the Conceptual Framework

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

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

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

Evaluating the PRO Instrument Hierarchy Using Recently Approved Labeling

The explicit relationships between the PRO instrument's conceptual content, expressed in terms of the PRO Concept Taxonomy, and the treatment benefit statements in labeling, reflected in the PRO Instrument Hierarchy, were evaluated and validated in two separate stages. A previous analysis showed that labeling for 64 (30%) of the 215 new drugs approved from 1997 to 2002 included a treatment benefit statement (in the Clinical Studies section) about a concept measured by a PRO instrument [20]. We first reanalyzed the labeling for these 64 drugs, and classified the conceptual frameworks represented by the PRO statements therein into one of the nine categories described in Table 1. During this first stage, the PRO Instrument Hierarchy was adapted to better fit the actual labeling statements observed. To validate this hierarchy, we then analyzed the labeling of the 142 new drugs approved in 2003–2007 (following the same criteria used in the 1997–2002 study), of which 36 contained PRO-based

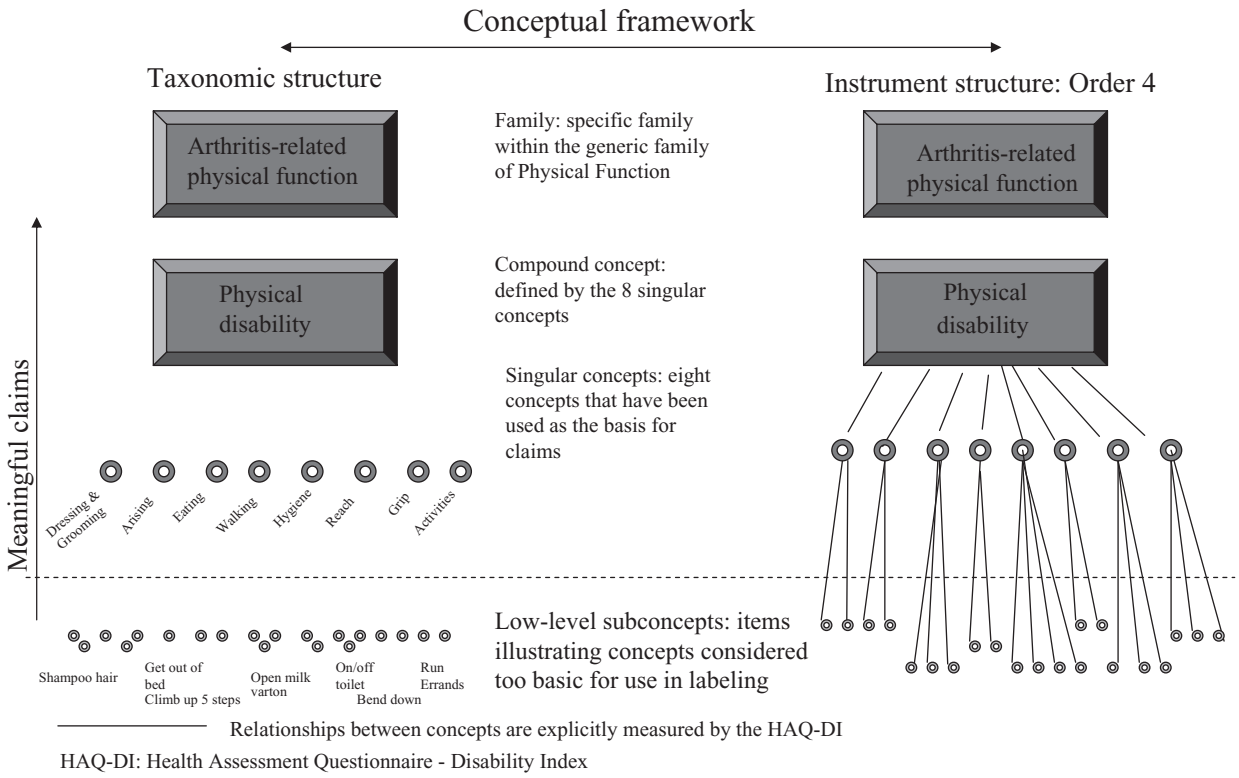
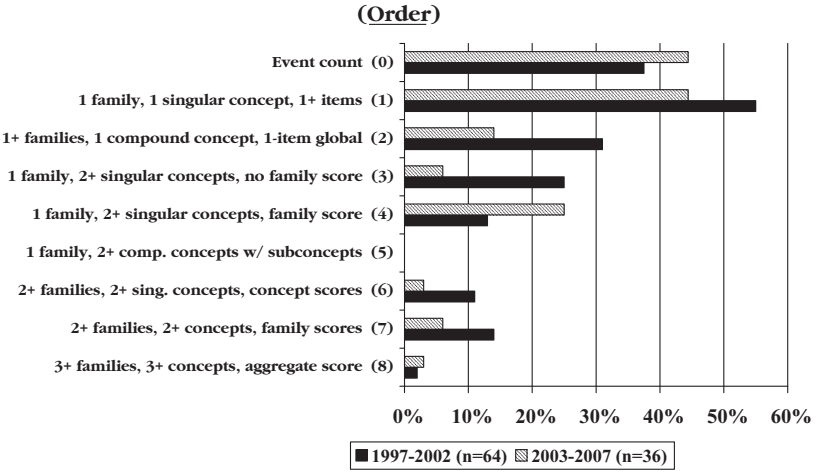


Figure 4 Intended statement of treatment benefit using the HAQ-DI: improve physical disability in rheumatoid arthritis patients.

statements in their Clinical Studies section, to determine whether those statements and their implied conceptual frameworks mapped well into the hierarchy. This second mapping determined that no changes to the basic structure of the hierarchy were needed, but we felt it was appropriate to modify the description of Order 4, from "There are at least 3 claimable end points" to "This may allow 3 or more claimable end points."

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

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



*Percentage of drugs with given or ders in their labeling, among all approved drugs with PROs. Many drugs have more than 1 order in their labeling.

Figure 5 Patient-reported outcome (PRO) instrument orders in new drug labeling*.

were cough (immunologic agents), partial seizure frequency (antiepileptic agents), and use of rescue medications (antimigraine and respiratory agents). Frequently used singular PRO concepts were: pain intensity, symptom assessments (several areas), ocular itching (ophthalmics), and dyspnea (cardiovascular).

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

Order 3 PRO measures (which measure a cluster of singular concepts) were most common among gastrointestinal agents and antimigraine products, where different symptom concepts (e.g., phonophobia, photophobia, nausea) are clustered together as a single disease-specific family of concepts (migraine symptoms). Use of Order 3 instruments was much higher in the 1997–2002 than in 2003–2007 due to the approval of 6 migraine drugs in the earlier period, all with Order 3 PRO measures, as opposed to no migraine drugs in the later period. In the earlier period, all but one of the approvals based on Order 4 measures referenced the HAQ Disability Index (or M-HAQ) for anti-inflammatory products [23,35]; the only other Order 4 instrument was the total nasal and non-nasal symptom score for a respiratory product. In the later period, however, there was more varied use of Order 4, including the Erectile Function domain of the International Index of Erectile Dysfunction, the Functional Living Index—Emesis, the Sheehan Disability Scale, and the Alzheimer's Disease Cooperative Study-Activities of Daily Living Inventory (ADCS-ADL) [36–39].

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

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

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

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

Discussion

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

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

The review of 1997–2002 new drug labeling illustrated that the PRO Instrument Hierarchy, incorporating the principles of the PRO Concept Taxonomy, is relevant across a wide range of both therapeutic products and the measures chosen to demonstrate their clinical benefit; this finding was confirmed by a subsequent review of 2003–2007 new drug labeling. For example, the predominance of the use of simple PRO instruments—event counts and singular concept PRO instruments (Orders 0 and 1)—along with global items and disease-specific, single-family PRO instruments (Orders 2 and 3) fits with the specific state-

ments about treatment benefit. Aside from the global PRO instruments, which are rarely used in isolation, the connection between the PRO instrument and the disease or its treatment is probably most transparent in these cases and the underlying conceptual framework of the instrument need not be complex.

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

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

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

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

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

Source of financial support: None.

References

1 US Food and Drug Administration. Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims, draft guidance. February 2006.
2 Bush JW. General health policy model/Quality of Well-Being (QWB) scale. In: Wenger NK, Mattson ME, Furberg CD, et al., eds., Assessment of Quality of Life in Clinical Trials of Cardiovascular Therapies. New York: New York, Le Jacq Publishing, 1984.
3 EuroQol Group. EuroQol—a new facility for the measurement of health-related quality of life. Health Policy 1990;16:199–208.
4 Ware JE, Sherbourne CD. The MOS 36-item Short Form Health Survey (SF-36): I. conceptual framework and item selection. Med Care 1992;30:473–83.
5 Katz S, Ford AB, Moskowitz RW, et al. Studies of illness in the aged: the index of ADL: a standardized measure of biological and psychosocial function. JAMA 1963;185:914–19.
6 OLGA. The on-line guide to quality-of-life assessment. Available from: http://www.OLGA-QoL.com [Accessed August 20, 2009].
7 McDowell I. Measuring Health: A Guide to Rating Scales and Questionnaires (3rd ed.). Oxford, England: Oxford University Press, 2006.
8 Kane RA, Kane RL, Eds. Assessing Older Persons: Measures, Meaning, and Practical Applications. Oxford, England: Oxford University Press, 2000.
9 Turk DC, Melzack R. Handbook of Pain Assessment (2nd ed.). New York, NY: The Guilford Press, 2001.
10 Bech P. Rating Scales for Psychopathology, Health Status and Quality of Life: A Compendium on Documentation in Accordance with the DSM-III-R and WHO Systems. Berlin, Germany: Springer-Verlag, 1993.
11 Frank-Stromberg M, Ed. Instruments for Clinical Health-Care Research (3rd ed.). Boston, MA: Jones and Bartlett Publishers, 2004.
12 Fries JF. The hierarchy of quality-of-life assessment, the Health Assessment Questionnaire (HAQ), and issues mandating development of a toxicity index. Control Clin Trials 1991;12(Suppl. 4):106S–17S.
13 Guyatt GH, Jaeschke R, Feeny DH, Patrick DL. Measurements in clinical trials: choosing the right approach. In: Spilker B, ed., Quality of Life and Pharmacoeconomics in Clinical Trials (2nd ed.). Philadelphia, PA: Lippincott-Raven Press, 1996.
14 Spilker B, Revicki DA. Taxonomy of quality of life. In: Spilker B, ed., Quality of Life and Pharmacoeconomics in Clinical Trials (2nd ed.). Philadelphia, PA: Lippincott-Raven Press, 1996.
15 American Psychiatric Association: Diagnostic and Statistical Manual of Mental Disorders (4th ed., Text Revision). Washington, DC: American Psychiatric Association, 2000.
16 International Statistical Classification of Diseases and Related Health Problems (10th Revision, Version for 2007). Available from: http://www.who.int/classifications/apps/icd/icd10online [Accessed August 20, 2009].
17 International Classification of Functioning, Disability, and Health. Available from: http://www.who.int/classifications/icf/en [Accessed August 20, 2009].
18 Verhoeff J, Toussaint PJ, Zwetsloot-Schonk JHM, et al. Effectiveness of the introduction of an International Classification of Functioning, Disability and Health-based Rehabilitation Tool in Multidisciplinary Team Care in patients with rheumatoid arthritis. Arthritis Rheum 2007;57:240–8.
19 Kennedy C. Functioning and disability associated with mental disorders: the evolution since ICIDH. Disabil Rehabil 2003;25:611–19.

20 Willke R, Burke LB, Erickson P. Measuring treatment impact: a review of patient-reported outcomes and other efficacy endpoints in approved labels. Control Clin Trials 2004;25:535–52.
21 Guidance for industry and review staff: target product profile—a strategic development process tool: draft guidance. Available from: http://www.fda.gov/downloads/Drug/Guidance/Compliance/RegulatoryInformation/Guidances/ucm080593 [Accessed August 20, 2009].
22 Erickson P, Scott J. Guide to Quality-of-Life Assessment (OLGA): a resource for selecting quality-of-life assessments. In: Walker S, Rosser R, eds., Quality of Life Assessment: Key Issues in the 1990s. Boston, MA: Kluwer Academic Publishers, 1993.
23 Fries JF, Spitz PW, Young DY. The dimensions of health outcomes: the Health Assessment Questionnaire, Disability and Pain Scales. J Rheumatol 1982;9:789–93.
24 Patrick DL, Erickson P. Health Status and Health Policy: Quality of Life in Health Care Evaluation and Resource Allocation. New York, NY: Oxford University Press, 1993.
25 Stewart AL. The medical outcomes study framework of health indicators. In: Stewart AL, Ware JE Jr, eds., Measuring Functioning and Well-Being: The Medical Outcomes Study Approach. Durham, NC: Duke University Press, 1992.
26 Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life: a conceptual model of patient outcomes. JAMA 1995;273:59–65.
27 Rothman ML, Beltran P, Cappelleri JC, et al. Patient-reported outcomes: conceptual issues. Value Health 2007;10(Suppl. 2):S66–75.
28 Patrick DL, Burke LB, Powers JH, et al. Patient-reported outcomes to support medical product labeling claims. Value Health 2007;10(Suppl. 2):S125–37.
29 Scientific Advisory Committee of the Medical Outcomes Trust. Assessing health status and quality-of-life instruments: attributes and review criteria. Qual Life Res 2002;11:193–205.
30 Blazeby J, Sprangers M, Cull A, et al. Guidelines for developing questionnaire modules. Available from: http://www.groups.eortc.be/qol/documentation_manuals.htm [Accessed August 20, 2009].
31 Leidy NK, Vernon M. Perspectives on patient-reported outcomes: content validity and qualitative research in a changing clinical trial environment. Pharmacoeconomics 2008;26:363–70.
32 Ware JE Jr, Sherbourne CD. The MOS 36-Item Short-Form Health Survey (SF-36): I. conceptual framework and item selection. Med Care 1992;30:473–83.
33 Felson DT, Anderson JJ, Boers M, et al. American College of Rheumatology. Preliminary definition of improvement in rheumatoid arthritis. Arthritis Rheum 1995;38:727–35.
34 Felson DT, Anderson JJ, Lange ML, et al. Should improvement in rheumatoid arthritis clinical trials be defined as fifty percent or seventy percent improvement in core set measures, rather than twenty percent? Arthritis Rheum 1998;41:1564–70.
35 Pincus T, Summey JA, Soraci SA Jr, et al. Assessment of patient satisfaction in activities of daily living using a modified Stanford Health Assessment Questionnaire. Arthritis Rheum 1983;26:1346–53.
36 Rosen RC, Riley A, Wagner G, et al. The International Index of Erectile Dysfunction (IIEF): a multidimensional scale for assessment of erectile dysfunction. Urology 1997;49:822–30.
37 Lindley CM, Hirsch JD, O'Neill CV, et al. Quality of life consequences of chemotherapy-induced emesis. Qual Life Res 1992;1:331–40.
38 Sheehan DV. The Anxiety Disease. New York: Scribner, 1986.
39 Galasko D, Bennett D, Sano M, et al. An inventory to assess activities of daily living for clinical trials in Alzheimer's disease. Alzheimer Dis Assoc Disord 1997;11(Suppl. 2):S33–9.
40 Beusterien KM, Rogers AE, Walsleben JA, et al. Health-related quality of life effects of modafinil for treatment of narcolepsy. Sleep 1999;22:757–65.
41 Comella CL, Stebbins GT, Goetz CG, et al. Teaching tape for the motor section of the Toronto Western Spasmodic Torticollis Scale. Mov Disord 1997;12:570–5.
42 Bellamy N, Buchanan WW, Goldsmith CH, et al. Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. J Rheumatol 1998;15:1833–40.
43 Burckhardt CS, Clark SR, Bennett RM. The fibromyalgia impact questionnaire: development and validation. J Rheumatol 1991;18:728–33.
44 Gilson BS, Gilson JS, Bergner M, et al. The sickness impact profile: development of an outcome measure of health care. Am J Public Health 1975;65:1304–10.
45 Aaronson NK, Ahmedza S, Bergman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality of life instrument for use in international clinical trials in oncology. J Natl Cancer Inst 1993;85:365–76.
46 Drugs@FDA. FDA approved drug products. Available from: http://www.accessdata.fda.gov/scripts/cder/drugsatfda/ [Accessed August 20, 2009].
47 PDR: Physicians' Desk Reference (Edition 61). Montvale, NJ: Medical Economics Company, 2007.

BMJ

BMJ 2013;346:f167 doi: 10.1136/bmj.f167 (Published 28 January 2013)

ANALYSIS

Patient reported outcome measures could help transform healthcare

Nick Black *professor of health services research*

London School of Hygiene and Tropical Medicine, London WC1H 9SH, UK

Abstract

Routine use of patient reported outcome measures (PROMs) has the potential to help transform healthcare, **says Nick Black**. Not only can PROMs help patients and clinicians make better decisions, but they can also enable comparisons of providers' performances to stimulate improvements in services

Patient reported outcome measures (PROMs) can drive the changes in how healthcare is organised and delivered. Key to this will be to link doctors' use of PROMs in the treatment of their patients with collection and aggregation of the data for assessing and comparing the performance of providers—all to improve healthcare quality.

What are PROMs?

Involvement of patients has moved on from simply seeking people's satisfaction with their care. PROMs seek to ascertain patients' views of their symptoms, their functional status, and their health related quality of life. PROMs are often wrongly referred to as so called "outcome measures," though they actually measure health—by comparing a patient's health at different times, the outcome of the care received can be determined. It's important to distinguish PROMs from patient reported experience measures (PREMs), which focus on aspects of the humanity of care, such as being treated with dignity or being kept waiting.

PROMs were initially developed for use in research, which has culminated in some regulatory bodies mandating their use. From there, PROMs were adopted by some doctors to enhance the clinical management of individual patients. In recent years they have been used to assess and compare the outcomes achieved by healthcare providers, with support of leading clinicians and encouragement of politicians. Some doctors still question their use, but most recognise the benefits of incorporating the views of patients (see box 1) alongside their own.

Current approaches to measuring patient reported outcomes

Broadly there are two types of PROM: disease specific and generic. The former, of which there are thousands, are tailored to the symptoms and impact on function of a specific condition. Generic PROMs consider general aspects such as self care and mobility (see box 2). Often both types are used, the former having greater face validity and credibility, the latter allowing comparisons across conditions. The reliability of PROMs is similar to that of clinical measures such as diastolic blood pressure or blood glucose.[1]

In addition to such multi-item PROMs, patients might also be asked single questions about the extent of any change in their health resulting from treatment (so called single transitional items) and also questions about any adverse consequences (complications).

Little is yet known about the impact of PROMs, although randomised trials of their use in clinical practice have demonstrated improvements in processes (such as diagnosis) and, less convincingly, in health outcomes.[2] The clearest benefits have been found in the diagnosis of depression. The more recent adoption of PROMs in comparing providers' performance means that their impact has not yet been evaluated.

How widely have PROMs been implemented in routine practice?

Individual clinicians and hospitals are increasingly using PROMs, but widespread use by health systems is still uncommon and largely restricted to England, Sweden, and parts of the United States. In contrast to England, where their adoption has been driven by government wishes for public comparisons of providers' performance, in Sweden and the US it has been the medical profession that has led the way, focused on improving the clinical care of individual patients.

In England, the principal use has been in elective surgery. The first nationwide application was in 2008 in a voluntary audit of mastectomy and breast reconstruction,[3] followed in April 2009 by a mandatory audit of all providers of hip and knee replacement, groin hernia repair, and varicose vein surgery (see box 3).[4] There are plans for more procedures to be added to this list, starting with coronary revascularisation in 2013. In addition, the feasibility of extension to long term conditions, cancer survivors, and people with dementia, is being explored.

Correspondence to: nick.black@lshtm.ac.uk

Box 1: Why consider patients' views?

- Most healthcare aims to reduce symptoms, minimise disability, and improve quality of life—these are aspects that only patients can assess
- Patients welcome being involved, and this may have health benefits in itself
- Patients' response rates are invariably better than clinicians' (a patient only has to complete one questionnaire whereas a clinician has to do it for every patient)
- The measure avoids observer bias (inevitable if asking clinicians to assess their own practice)
- Considering patients' views increases public accountability of health services and healthcare professionals

Box 2: Example of a disease specific and a generic PROM

Disease specific PROM: Oxford Hip Score

Twelve questions about how the patient has been over the previous 4 weeks covering pain (4 items), mobility (3 items), and activities (5 items). Five possible answers scored from 0 to 4, creating overall scale of 0 (severe disease) to 48 (no problems).

Example questions:

- During the past 4 weeks have you been able to climb a flight of stairs?
- Yes, easily/With little difficulty/With moderate difficulty/With extreme difficulty/No, impossible
- During the past 4 weeks how would you describe the pain you usually had from your hip? None/Very mild/Mild/Moderate/Severe
- During the past 4 weeks could you do the household shopping on your own?
- Yes, easily/With little difficulty/With moderate difficulty/With extreme difficulty/No, impossible

Generic PROM: EuroQol EQ-5D

Five questions seeking information that best describes the patient's health that day, covering mobility, self care, usual activities, pain/discomfort, anxiety/depression. Three possible answers: no problem; some problem; severe problem.

Example questions:

Self care: I have no problems with self care/I have some problems washing or dressing myself/I am unable to wash or dress myself.

Anxiety/depression: I am not/moderately/extremely anxious or depressed.

Given that there are over 50 established national clinical audits (all but one limited to clinicians' reports of processes and outcomes), opportunities for wider use of PROMs are readily available.

Nationwide use of PROMs commenced earlier in Sweden using the disease specific clinical databases (quality registers) established there by the medical profession since 1975.[5] PROMs began to be introduced in some in 2000.[6] In the US, widespread implementation of PROMs has been more restricted: for spinal conditions in northern New England,[7] for primary care in Pittsburgh,[8] and for depression in Minnesota.[9] The only nationwide use has been to compare health plans that purchase care for those over 65 years of age (Medicare).[10] In 2015, the federal government plans to extend the use of PROMs to reimbursement mechanisms for accountable care organisations (health maintenance organisations with a focus on outcome measurement). It is hoped that this will enable the level of reimbursement to reflect the value that patients' ascribe to the outcome of their treatment.[8]

How are PROMs being used in England?

All three ways that PROMs can improve care are being pursued in England: assisting clinicians to provide better and more patient centred care; assessing and comparing the quality of providers; and providing data for evaluating practices and policies.

As regards the first, PROMs are being used to monitor patients' conditions to help them and their doctors make well informed decisions about their treatment.[11] For example, three monthly measurements by people with hip osteoarthritis to help clinicians decide if and when to operate.[12] Similarly, regular reporting of PROMs is being used to help patients and doctors share the management of long term medical conditions.[13] PROMs help clinical decision making in the same way clinical investigations do. They are not used as absolute determinants ("patients with an Oxford Hip Score under 30 should have surgery," for example) because their predictive validity for individuals is not strong enough.

The second use, for provider comparisons, aims to stimulate improvements in quality in several ways.

Firstly, patients can choose where to be treated on the basis of the outcome reports of other patients, though in practice many other factors (such as distance from home) also influence a patient's preference.[14] Secondly, PROMs are included in the NHS outcomes framework, which will be used to performance manage the NHS Commissioning Board and clinical commissioning groups. Thirdly, by having to report PROMs in their annual quality account, NHS providers account publicly for their performance to their local community. Fourthly, PROMs data can contribute to the revalidation of doctors. Finally, PROMs provide a means of enhancing the calculation of healthcare productivity by including the outcome as well as the quantity of care.

As for research, routine PROMs provide data on large numbers of patients representative of typical, everyday practice, thus facilitating research on the effectiveness (rather than efficacy) of treatments.[15] The inclusion of generic PROMs (such as the EuroQol EQ-5D—see box 2) allow patient utilities to be derived for cost effectiveness analysis. Such data can also be used to evaluate policies quickly and cheaply, such as the introduction of new ways of providing care[16] and the equity of services.[17]

What are the challenges and how can they be met?

Despite good progress in introducing PROMs into routine practice, more widespread implementation faces several challenges.

Box 3: National PROMs programme in England for elective surgery

From April 2009 it has been mandatory for all providers (NHS hospitals, independent sector treatment centres, private hospitals) treating NHS patients for any of four elective procedures to participate in the national PROMs programme. All patients undergoing a hip or knee replacement, groin hernia repair, or varicose vein surgery should be invited to complete a questionnaire before surgery, either at the pre-assessment clinic or on the day of admission.

The preoperative questionnaire collects data on the patient's sociodemographic characteristics, the duration of their condition, their general health, any comorbidities, and whether they are undergoing a repeat/revision procedure. In addition, they are asked to complete a disease specific PROM (Oxford Hip Score, Oxford Knee Score, or Aberdeen Varicose Vein Score; there is no available instrument for hernia repair) and a generic PROM (EQ-5D index and EQ-Visual Analogue Scale).

Patients who complete a preoperative questionnaire are mailed a postoperative questionnaire after three months (hernia repair, varicose vein surgery) or six months (hip or knee replacement). Non-responders receive one reminder letter. The questionnaire includes the same PROMs as the preoperative one plus single transitional items on their overall view of the result of surgery and the extent of any improvement. They are also asked to report on adverse outcomes (complications, readmission, and further surgery).

Over the first two years, of the 485 000 eligible patients, 329 000 (68%) were recruited, though this varied from about 80% recruitment for hip and knee replacement to 60% for hernia repair and 50% for varicose vein surgery. Postoperative response rates also differed by procedure from 85% (hip and knee replacement) to 75% (hernia repair) and 65% (varicose vein surgery).

PROMs data are linked to Hospital Episode Statistics by the Health and Social Care Information Centre who provide regular analysis of each provider's preoperative patient characteristics (age, sex, severity) and the mean change in the PROM scores adjusted for case mix.[4] Providers are identified and compared by means of funnel plots that show whether or not any provider's outcome is significantly different from what would be expected.

Minimising the time and cost of collection, analysis, and presentation of data

Information technology is important. Patient reported measurement systems are already being developed and web based entry has been introduced not only in clinical settings but also in patients' homes.[5] [7] [12] [13] However, implementation is not necessarily straightforward. For example, although rheumatology departments in Sweden started converting to web based entry in 2003, by 2012 only 39% had done so. Hip replacement patients are more likely to respond to mailed (92%) than internet based questionnaires (49%).[18] Another option for minimising cost is to reduce the number of data collected by replacing multi-item PROMs with single transitional items.[19] This approach is the basis of the current quest by the Department of Health in England for a short questionnaire that could be used for all conditions and interventions.

Achieving high rates of patient participation

The challenge is how to achieve high rates particularly among older, sicker, more deprived, and non-white patients who tend to be under represented.[20] It is harder to recruit patients with minor conditions or those undergoing minor (or no) procedures, and those who are outpatients rather than inpatients.[21] As for primary care, little is yet known but it is likely to require different, innovative approaches, particularly for repeated assessments of patients with long term conditions.

Recognising all three dimensions of quality: safety, effectiveness, experience

PROMs focus mostly on the effectiveness of care, but safety and experience, the other two key dimensions of quality, must not be ignored. It is known that poor safety (such as complications) has an adverse impact on patient's perception of the effectiveness of care.[22] The impact of patients' experience of the humanity of care (such as dignity and respect) has started to be considered but requires much more investigation. It may be that judgments of a provider's effectiveness will need to be adjusted to take into account patients' experiences and vice versa.

Attributing outcomes to the quality of care

This presents several challenges.

Firstly, meaningful comparisons of providers require sufficiently robust adjustment for differences in case mix to achieve credibility. In addition to collecting data on known confounders from patients, more use could be made of obtaining data through linkage with other databases.

Secondly, judging the best time to assess outcome after an intervention so as to be able to attribute it to that intervention is often contentious: delaying follow-up ensures patients have gained all possible benefit but may undermine attribution to the intervention in question.

The third issue is determining the appropriate level of analysis for attributing responsibility for a patient's outcome. Currently most PROMs are reported at institutional level (such as that of hospital, trust, commissioner). While this may be appropriate for some interventions, for others the individual practitioner is perceived as the attributable level. This is true of surgery—patients and surgeons (and many politicians) are enthusiastic for data at this level.[14] In contrast, the treatment of long term conditions depends on both primary and secondary care, so whole health economies may be the appropriate entity to consider.

Fourthly, emergency admissions present a challenge in the attribution of impact of care when PROMs are only available after the event. Solutions that need exploring are a patient's recall of their pre-event health and the use of population norms.

Providing appropriate output to different audiences

Most questionnaires include more than one PROM (for example, a disease specific and a generic measure), each of which may draw different conclusions about a provider's performance. In addition, different metrics can be derived from a measure (such as the mean PROM score or the proportion of patients achieving a certain level of improvement), and these may also assess providers' performances differently. Further, having decided on an indicator, defining what constitutes unacceptable performance requires careful consideration (fig 1⇓). The comparative risks of missing a poor performer must be weighed against unfairly criticising a provider. It is unclear if the rules used for clinical outcomes (such as mortality) are appropriate for PROMs. Is a PROM score more than three standard deviations from the mean as serious as a death rate that far from the mean? Also, deciding how to present the data needs to be tailored to the intended audience.[23] [24]

Avoiding misuse of PROMs

There is a danger of PROMs being used crudely to ration care. Data from the national PROMs programme have already been misinterpreted as showing that 20 000 hernia and varicose vein

operations and 16 000 hip and knee replacements each year should not take place.[25] While some patients will not benefit from surgery, unfortunately they cannot necessarily be identified preoperatively using PROMs. Another potential misuse is using PROMs to decide on competing demands between treatments for funding. If only short term outcomes are considered and longer term aspects, such as the natural history of the disease or long term outcomes, are ignored, poor conclusions may be drawn.[26]

Where next for PROMs?

The routine use of PROMs provides an opportunity to help drive changes in how healthcare is organised and delivered. There are five priorities for maximising their contribution.

Firstly, despite their separate development to date, we need to combine initiatives to use PROMs for clinical management and for provider comparisons, to contribute to both goals. Secondly, we need to encourage the adoption of new data collection technologies such that PROMs become part of everyday care. Thirdly, given that it is not feasible to extend provider comparisons to all healthcare, priority diseases and treatments need to be identified. Fourthly, we need to tackle the methodological challenges that remain unresolved to ensure PROMs are used appropriately. And finally, we must make use of the opportunity that PROMs presents to develop value based care in which health services can be driven by health outcomes per pound spent.[27]

Contributor: NB has played a leading role in the development and establishment of the NHS national PROMs programme in England. As a grant holder for methodological studies of PROMs from the Department of Health Policy Research Programme, he was a member of the DH PROMs Operations Board. He is also a member of the DH PROMs Stakeholder Reference Group and chairs the DH National Advisory Group for Clinical Audit and Enquiries. NB is sole contributor and guarantor.

Competing interests: The author has completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declares: no support from any organisation for the submitted work; a financial relationship with the Department of Health for research on PROMs and membership of the DH Operations Board for the National PROMs Programme.

Provenance and peer review: Commissioned; externally peer reviewed.

1 Hahn EA, Cella D, Chassany O, Fairclough DL, Wong GY, Hays RD; Clinical Significance Consensus Meeting Group. Precision of health-related quality-of-life data compared with other clinical measures. *Mayo Clin Proc* 2007;82:1244-54.
2 Valderas JM, Kotzeva A, Espallargues M, Guyatt G, Ferrans CE, Halyard MY, et al. The impact of measuring patient-reported outcomes in clinical practice: a systematic review of the literature. *Qual Life Res* 2008;17:179-93.
3 Jeevan R, Cromwell D, Browne J, van der Meulen J, Pereira J, Caddy C, et al. Fourth Annual National Mastectomy and Breast Reconstruction Audit 2011. The NHS Information Centre, 2011.
4 NHS Information Centre. National Patient Reported Outcomes Programme. www.ic.nhs.uk/proms.
5 Swedish Association of Local Authorities and Regions. Quality registries. www.kvalitetsregister.se/om_kvalitetsregister/quality_registries.
6 Rolfson O, Karrholm J, Dahlberg LE, Garellick G. Patient-reported outcomes in the Swedish Hip Arthroplasty Register: results of a nationwide prospective observational study. *J Bone Joint Surg (Br)* 2011;93:867-75.
7 Nelson EC. Using patient-reported information to improve health outcomes and health care value: case studies from Dartmouth, Karolinska and Group Health. Dartmouth Institute for Health Policy and Clinical Practice, June 2012.
8 Hostetter M, Klein S. Using patient-reported outcomes to improve health care quality. *Quality Matters*. January 2012, The Commonwealth Fund. www.commonwealthfund.org/Newsletters/Quality-Matters/2011/December-January-2012/In-Focus.aspx.
9 Minnesota Community Measurement. Minnesota Health Scores Overview. www.mnhealthscores.org.
10 Centers for Medicare and Medicaid Services (CMS). Medicare health outcomes survey. www.hosonline.org/Content/Default.aspx.
11 Stiggelbout AM, Van der Weijden T, De Wit MP, Frosch D, Légaré F, Montori VM, et al. Shared decision making: really putting patients at the centre of healthcare. *BMJ* 2012;344:e256.
12 My clinical outcomes. Helps patients and doctors manage long term conditions. www.myclinicaloutcomes.co.uk.
13 PROMs 2.0. http://proms2.org/default.html.
14 Coulter A. Do patients want a choice and does it work? *BMJ* 2010;341:c4989.
15 Baker PN, Petheram T, Jameson SS, Avery PJ, Reed MR, Gregg PJ, et al. Comparison of patient-reported outcome measures following total and unicondylar knee replacement. *J Bone Joint Surg Br* 2012;94:919-27.
16 Chard J, Kuczawski M, Black N, van der Meulen J. Outcomes of elective surgery undertaken in independent sector treatment centres and NHS providers in England: audit of patient outcomes in surgery. *BMJ* 2011;343:d6404.
17 Neuburger J, Hutchings A, Allwood D, Black N, van der Meulen JH. Sociodemographic differences in the severity and duration of disease in patients undergoing hip or knee replacement surgery. *J Public Health (Oxf)* 2012;34:421-9.
18 Rolfson O, Salomonsson R, Dahlberg LE, Garellick G. Internet-based follow-up questionnaire for measuring patient-reported outcome after total hip replacement surgery—reliability and response rate. *Value Health* 2011;14:316-21.
19 Grosse Frie K, van der Meulen J, Black N. Single item on patients' satisfaction with condition provided additional insight into impact of surgery. *J Clin Epidemiol* 2012;65:619-26.
20 Hutchings A, Grosse Frie K, Neuburger J, van der Meulen J, Black N. Late response to patient-reported outcome questionnaires after surgery was associated with worse outcome. *J Clin Epidemiol* 2013;66:218-25.
21 Royal College of Obstetricians and Gynaecologists. National Heavy Menstrual Bleeding Audit. Second Report. July 2012. www.rcog.org.uk/files/rcog-corp/NationalHMBAudit_2ndAnnualReport_11.07.12_forweb.pdf.
22 Grosse Frie K, van der Meulen J, Black N. Relationship between patients' reports of complications and symptoms, disability and quality of life after surgery. *Brit J Surg* 2012;99:1156-63.
23 Allwood D, Hildon Z, Black N. Clinicians' views of formats of performance comparisons. *J Eval Clin Pract* 2011; doi:10.1111/j.1365-2753.2011.01777.x.
24 Hildon Z, Allwood D, Black N. Making data more meaningful. Patients' views of the format and content of quality indicators comparing health care providers. *Patient Educ Couns* 2012;88:298-304.
25 West D. Unneeded surgery may be costing the NHS millions. *Health Services Journal*. 14 May 2009. www.hsj.co.uk/news/acute-care/unneeded-surgery-may-be-costing-the-nhs-millions/5001423.article.
26 Smith PC, Street AD. On the uses of routine patient-reported health outcome data. *Health Econ* 2013;22:119-31.
27 Porter M. A strategy for health care reform—toward a value-based system. *N Engl J Med* 2009;361:109-12.

Cite this as: *BMJ* 2013;346:f167

Key messages

- PROMs can be used to support the clinical management of patients, assess provider performance, and provide a basis for evaluative research
- Nationwide use is most advanced in England (particularly for performance comparisons) and Sweden (for supporting clinical practice)
- Several challenges need to be addressed including minimising costs, achieving high participation, attributing causality, providing appropriate outputs and discouraging misuse of PROMs data
- The separate development of PROMs in clinical practice and in provider comparisons needs to be brought together for the benefit of both tasks
- The impact of PROMs on clinical practice and on stimulating improvements in the quality of health services still needs to be established

Figure

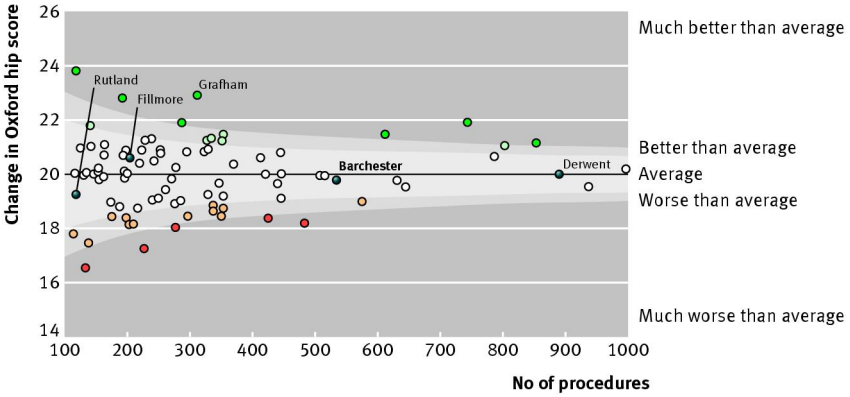


Fig 1 Funnel plot of mean change in Oxford Hip Score following primary hip replacement for 88 NHS trusts (real data; fictitious names). Note that trusts more than three standard deviations below average ("much worse than average") have mean scores only 2-3 points below average

Health and Quality of Life Outcomes

Commentary

The FDA guidance for industry on PROs: the point of view of a pharmaceutical company

Fabio Arpinelli* and Francesco Bamfi

Address: Health Technology Assessment, Medical Department, GSK S.p.A. Verona, Italy

Email: Fabio Arpinelli* - fabio.a.arpinelli@gsk.com; Francesco Bamfi - francesco.a.bamfi@gsk.com

* Corresponding author

Published: 31 October 2006

Health and Quality of Life Outcomes 2006, **4**:85 doi:10.1186/1477-7525-4-85

Received: 03 October 2006

Accepted: 31 October 2006

This article is available from: http://www.hqlo.com/content/4/1/85

Abstract

The importance of the patients point of view on their health status is widely recognised. Patient-reported outcomes is a broad term encompassing a large variety of different health data reported by patients, as symptoms, functional status, Quality of Life and Health-Related Quality of Life. Measurements of Health-Related Quality of Life have been developed during many years of researches, and a lot of validated questionnaires exist. However, few attempts have been made to standardise the evaluation of instruments characteristics, no recommendations are made about interpretation on Health-Related Quality of Life results, especially regarding the clinical significance of a change leading a therapeutic approach. Moreover, the true value of Health-Related Quality of Life evaluations in clinical trials has not yet been completely defined. An important step towards a more structured and frequent use of Patient-Reported Outcomes in drug development is represented by the FDA Guidance, issued on February 2006.

In our paper we aim to report some considerations on this Guidance. Our comments focus especially on the characteristics of instruments to use, the Minimal Important Difference, and the methods to calculate it. Furthermore, we present the advantages and opportunities of using the Patient-Reported Outcomes in drug development, as seen by a pharmaceutical company. The Patient-Reported Outcomes can provide additional data to make a drug more competitive than others of the same pharmacological class, and a well demonstrated positive impact on the patient' health status and daily life might allow a higher price and/or the inclusion in a reimbursement list. Applying extensively the FDA Guidance in the next trials could lead to a wider culture of subjective measurement, and to a greater consideration for the patient's opinions on his/her care. Moreover, prescribing doctors and payers could benefit from subjective information to better define the value of drugs.

Introduction

The importance of the patients point of view on their health status and healthcare is widely recognized [1]. Patient-reported outcomes (PROs) provide the patient's perspective on health outcome endpoint data [2-4]. PROs can play an important role in the development of new drugs, especially those aimed to treat medical conditions in which only subjective data allow to evaluate the treatment effect [1].

PROs is a broad term encompassing a large variety of different health data reported by patients. PROs as symp-

toms, functional status, treatment adherence, satisfaction with care represent useful data to corroborate the clinical data (efficacy and safety), helping clinicians to better define the drug profile.

Furthermore, inside the PROs we meet a couple of important concepts, sometimes considered as synonymous. These concepts are the Quality of Life (QoL) and the Health-Related Quality of Life (HRQoL). The QoL is a complex, abstract, multidimensional concept defining an individual satisfaction with life in domains he/she considers important. The HRQoL reflects an attempt to restrict the complex concept of QoL to those aspects of life specifically related to the individual health, and potentially modified by healthcare [5]. HRQoL data are not always foreseeable and necessarily correlated with the severity of the disease as perceived by healthcare professionals. Moreover, the symptoms/HRQoL correlation could be weak (for example, no abdominal pain during a medical examination but a poor patient's HRQoL, because of the impairment of the patient personal life and leisure, his/her need to take drugs, dietary restrictions etc.).

HRQoL measurements has been developed during many years of research (proven by thousand of published papers), and a lot of validated questionnaires exist, both generic and disease specific. However, the following points need to be considered: 1) although the operational application of concepts and their validation process have been well codified, few attempts have been made to standardise the evaluation of instruments characteristics; 2) usually, the criteria regard intrinsic characteristics of the questionnaires (reliability, validity etc.), while no recommendations are made about interpretation on HRQoL results, especially regarding the clinical significance of a change in HRQoL leading a therapeutic approach; 3) despite some scientific society have created working groups to debate the role of HRQoL in clinical research, the true value of HRQoL evaluations in clinical trials has not yet been completely defined [5,8,9].

The contribution given by the PROs measurement could be important in the process of drug approval by regulatory authorities. Furthermore, on the regulatory side some factors limit the use of PROs, and HRQoL in particular. The main limiting factors are: 1) the abuse of the term HRQoL in clinical trials. This term is used also when other PROs are measured (symptoms, drug side effects etc.). 2) The poor quality of the majority of clinical trials having the HRQoL as primary endpoints. 3) The role and the significance of HRQoL as efficacy, tolerance, utility endpoint [10,11]

These points and the reasonable scepticism of regulatory authorities to officially acknowledge some subjective criterion whose clinical meaning remains difficult, have limited the use of PROs in the drugs approval process. At the moment, it could be quite difficult to make acceptable HRQoL to regulatory authorities as a primary endpoint, since some regulators consider it as a less rigorous secondary endpoint.

An important step towards a more structured and frequent use of PROs in drug development has been done by FDA. On February 2006 the FDA issued the Guidance, that describes how it evaluates PROs used as effectiveness endpoints in clinical trials.

Specific comments to the Guidance

The Guidance is potentially very useful for all concerned in planning, designing and carrying out clinical trials for regulatory purposes. It provides information on how to choose a PRO instrument. Although the Guidance is clear enough and take into consideration a lot of important topics on PRO instruments (their development, assessment of measurement properties, modification of existing instruments), study design and data analysis, it could be improved to make it more applicable to NDA trials and facilitate univocal interpretation of results by experts and regulators.

The reading of the FDA Guidance firstly led to some general considerations and comments on the PROs. We aim to briefly report these considerations.

PROs is an "umbrella term". It contains physical functioning, psychological well-being, global health perception, treatment satisfaction and other subjective outcomes. Therefore, PRO is not interchangeable with QoL or HRQoL.

QoL has never been approved in a labelling claim because of its vagueness. On the contrary, HRQoL could be a possible endpoint. This should be very clear when measuring PROs.

The inclusion of PROs assessment in clinical trials should have a good scientific rationale. The risk of an indiscriminate measuring of PROs is producing useless and confounding data.

The conceptual framework of a single-item symptom measure is not so complex as a multiple frameworks to define HRQoL. Multiple domains questionnaires usually are required in early phases of drug development, when researchers investigate the activity of a new compound more than its efficacy. In this phase the need to focus the attention on the domains more affected by the disease (or by the disease management) makes useful a multi-domains questionnaire. In later phases, when needs and

expectations of pts are well known, a single or few domains questionnaire helps to a better interpretation of changes.

HRQoL measurements is more useful in chronic diseases (for example rheumatoid arthritis) than in life-threatening disease (cancer). In life-threatening diseases the only acceptable main aim of the therapy is a longer survival. A better HRQoL and a worsened survival make that drug probably not approved by regulatory authorities.

PROs should not replace safety reporting, as safety is an important concern of regulatory authorities.

Researchers aiming to measure HRQoL must use existing, validated instruments. The development of new questionnaires should be discouraged, but the standardisation of questionnaires should be encouraged. In particular, the development of a questionnaire for a certain study should be definitely avoided. The sponsor of the study has to provide evidence of validity of the selected instrument (for example, a list of published papers on the development, validation and use in clinical trials of the questionnaire).

When an existing questionnaire is used in a new population (elderly rather than adults) or in a different context (on outpatients basis rather than inpatients), a re-validation is required.

The instrument used to measure a PRO should have a documented evidence of responsiveness/sensitivity to changes in health status. In fact, small differences in PRO scores, although statistically significant, are often questioned with regard to their clinical importance. It is not always clear what is meant by clinically importance, i.e., discernible to the patient, significant enough for a clinician to change an intervention, or significant from a population perspective. Hence if a PRO cannot detect a meaningful change in health status, its use may be risky, because clinically meaningful effects may be undetected [2,6].

Demonstrating responsiveness is necessary to determine the Minimal Important Difference (MID), where MID represents the smallest change perceived by the patient as an advantage, or that could lead to a change of treatment [6].

The MID can be calculated using a number of anchor-based or distribution-based methods. Distribution-based approaches are the effect size (ES), the standardised response mean and the standard error of measurement (SEM). Anchor-based methods assess which changes on the measurement instrument correspond with a minimal important change defined on a anchor. Distribution-based methods do not provide a good indication of the importance of the observed change. Anchor-based approaches do not take measurement precision into account. Sometimes results obtained using these different approaches are similar [12].

At the moment, there is not a clear agreement on the recommended, best practice approach for determining the MID [7]. The application of multiple methods, even if imperfect, to the same datasets could tend to give similar results and this should clarify the relationship between these methods and give a better estimate of the MID. Furthermore, some Authors report that for assessing the MID anchor-based methods are preferred, as they include a definition of what is minimally important [12].

These concepts should be more stressed in the Guidance. The MID may vary by context, and different MID could be valid for different studies where PROs instruments are used. MID varies according different factors, such as the underlying disease, the characteristics of the population, the healthcare scenario, and so on. For these reasons, we cannot have a unique MID for a PRO instrument, good for different diseases and patients [7]. It is necessary that responsiveness and MID be well documented in order to use PROs in labelling claims.

The patient satisfaction is a PRO, but it could be greatly influenced by factors such as, for example, the personal relationship between the patient and the nurse/doctor. This relationship can satisfy/dissatisfy the patient, and represents an aspect related to the (variable) healthcare structure/organization. For this reason we believe that the patient satisfaction should be considered as a less important indicator than HRQoL.

The users of the Guidance should appreciate more details for sample size determination and handling missing data, especially for the questionnaire development. Another topic to be detailed is concerned with the proxy measures.

Furthermore, we are aware that the heterogeneity of clinical settings, diseases and drugs makes very difficult to anybody (including FDA) to prepare a technical documentation applicable to any context.

The point of view of a pharmaceutical company

Certainly the drug developers are interested in a better definition of the value of their drugs using PROs data. The pharmaceutical industry has been the principal driving force behind the expansion in the number and type of HRQoL instruments available to clinician and researchers [13].

Industry sees some advantages in PRO measurements. Infact, PROs can provide additional data for inclusion of

a drug in a formulary, making that drug more competitive than others of the same pharmacological class. Furthermore, an effective and well tolerated drug with a demonstrated positive impact on the patient' health status and daily life might allow to negotiate a higher price (where the price of drugs is negotiated between pharmaceutical companies and regulatory authorities) and/or the inclusion in a reimbursement list.

Subjective data collection has to be regulated by clear rules, agreed by all parts involved in the development and approval of drugs. The Guidance is a positive and modern attempt to provide a document helping the use of subjective data to support labelling claims. It stimulates pharmaceutical companies to use a shared and accepted methodology to provide data, although an alternative approach is considered possible. This means longer time to prepare and carry out a clinical trial, and more expenses. On the other side, the adherence to the Guidance should reduce the risk of rejection of PROs data by FDA.

Furthermore, the Guidance offers a good opportunity to the pharmaceutical industry to discuss about methodology with regulatory authorities, and to become a trustworthy partner of regulatory agencies.

The development of a new questionnaire (if needed) or their revision/updating is a complex, time consuming and expensive activity. Usually a pharmaceutical company can provide financial support and technical knowledge to develop subjective questionnaires by itself or in partnership with a scientific society and academic experts. Adequate resources can allow tool developers to reach an exhaustive set of data to demonstrate the validity and reliability of questionnaires. A further important step should be the publication of papers, allowing the developers to insert the new tools in a compendium, where all concerned researchers and regulators can find the instruments and replicate experiences to confirm the validity of the instruments.

Regulatory authorities might recognise that the development of a new instrument allows clinicians to have a useful instrument to administer to their patients. The companies could waive the copyright in favour of all researchers and clinicians, obtaining, on the other hand, both an increase of robustness of the tool and, maybe, a reward by regulators.

A wider use of PRO measurements allows clinicians/payers to become familiar with PROs, integrating these data in their evaluation criteria to prescribe or reimburse a drug.

Conclusion

In conclusion, it is well known that the correlation between the patient and the physician evaluation of a certain symptom could be poor and not univocal. HRQoL and other PROs provide important patient perspective on disease and the treatment they receive. A subjective evaluation provides clinically important information not captured by objective measures. This is particularly important in chronic diseases, as rheumatoid arthritis or asthma, where HRQoL data capture the overall benefit given by the treatment.

Despite a very large number of published papers on HRQoL, there is a certain scepticism on the value of HRQoL and other PROs. It is likely that clinicians do not use PROs because they are not routinely trained in the use and interpretation of PRO instruments [1]. Usually, the interpretation of the clinical significance of a change in HRQoL is considered difficult; particularly difficult is the translation of results into an overall clinical evaluation leading to a change of the current therapy [14,15].

In order to overcome this scepticism, it is necessary to highlight the scientific and statistical basis of these measurements, and the usefulness of collecting these data. Moreover, it shall be demonstrated the improvement of patients management by clinicians thanks to the use of PROs data.

The use of these data to support a labelling claim requires the use of a rigorous methodology, based on valid and reliable instruments, used when appropriate.

A parallel European Guidance has not yet been conceived by EMEA, and this is not surprising considering the differences between the American and European healthcare and regulatory structures. This is reflected in a different marketing approval process, which is first centrally granted (EMEA), and subsequently discussed at the national level (reimbursement, price). Furthermore, EMEA prepared a Reflection Paper (July 2005), a short and generic document that discusses the place that HRQoL may have in drug evaluation process, and gives some broad recommendations.

The FDA guidance represents the first step in a hard, complex track to reach the best evidence in questionnaire development and the use of PRO to support labelling claims.

Applying extensively the Guidance in the next trials could lead to a wider culture of subjective measurement, and to take into a greater consideration the patient's point of view on his/her care. Moreover, a more detailed evalua-

tion of drugs is helpful for prescribing doctors and payers, to allow them to better define the value of drugs.

Competing interests

The author(s) declare that they have no competing interests.

Authors' contributions

FA discussed the FDA guidance, exposed the point of view of a pharmaceutical company and drafted the manuscript. FB reviewed the manuscript in order to improve the parts that deal with statistics.

Acknowledgements

We would like to acknowledge Giovanni Apolone MD for his invaluable contribution in the critical reviewing of the draft of the manuscript.

References

1. Marquis P, Arnould B, Acquadro C, Roberts WM: **Patient-reported outcomes and health-related quality of life in effectiveness studies: pros and cons.** *Drug Development Research* 2006, **67:**193-201.
2. Leidy NK, Revicki DA, Geneste B: **Recommendations for evaluating the validità of qualità of life claims for labelling and promotion.** *Value Health* 1999, **2:**113-127.
3. Revicki DA, Osoba D, Fairclough D, Barofsky I, Berzon R, Leidy NK, Rothman M: **Recommendations on health-related quality of life research to support labelling and promotional claims in the United States.** *Qual Life Res* 2000, **9:**887-900.
4. Wilke RJ, Burke LB, Erickson P: **Measuring treatment impact a review of patient-reported outcomes and other efficacy endpoints in approved product labels.** *Contr Clin Trials* 2004, **25:**535-552.
5. Apolone G, De Carli G, Brunetti M, Garattini S: **Health-Related Qualità of Life and Regulatory Issues.** *Pharmacoeconomics* 2001, **19:**187-195.
6. Guyatt G, Walter S, Norman G: **Measuring change over time: assessing the usefulness of evaluative instruments.** *J Chronic Dis* 1987, **40:**171-178.
7. Guyatt G, Osoba D, Wu AW, Wyrwich KW, Norman GR: **Methods to explain the clinical significance of health status measures.** *Mayo Clinic Proceed* 2002, **77:**371-383.
8. Leplege A, Hunt S: **The problem of quality of life in medicine.** *JAMA* 1997, **278:**47-50.
9. Guyatt G, Feeny D, Patrick D: **Issues in quality of life measurement in clinical trials.** *Control Clin Trials* 1991, **12:**81S-90S.
10. Acquadro C, Berzon R, Dubois D, Leidy NK, Marquis P, Revicki D, Rothman M, PRO Harmonization Group: **Incorporating the patient's perspective into drug development and communication: an ad hoc task force report of the Patient-Reported Outcomes (PRO) Harmonization Group meeting at the Food and Drug Administration, February 16, 2001.** *Value Health* 2003, **6:**522-531.
11. Szende A, Leidy NK, Revicki D: **Health-related quality of life and other patient-reported outcomes in the European centralized drug regulatory process: a review of guidance documents and performed authorizations of medicinal products 1995 to 2003.** *Value Health* 2005, **8:**534-548.
12. De Vet HC, Terwee CB, Ostelo RW, Beckerman H, Knol DL, Bouter LM: **Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change.** *Health Qual Life Outcomes* 2006, **4:**54.
13. Delate T, Ernst FR, Coons SJ: **The role of Health-Related Quality of Life data in pharmacy benefit decision.** *P&T* 2002, **27:**24-32.
14. Donaldson MS: **Taking stock of health-related quality of life measurement in oncology practice in the United States.** *J Natl Cancer Inst Monogr* 2004, **33:**155-167.
15. Greenhalgh J, Long AF, Flynn R: **The use of patient reported outcome measures in routine clinical practice: lack of impact or lack of theory?** *Soc Sci Med* 2005, **60:**833-843.

Contents

Guidelines for assessing Quality of Life in EORTC clinical trials

1. AIMS OF THE QUALITY OF LIFE GROUP

1. To develop reliable and valid instruments for measuring the quality of life of cancer patients participating in international clinical trials.
2. To advise the EORTC about the assessment of the multidimensional aspects of patients' quality of life as a measurable outcome of cancer treatment. Where appropriate the Quality of Life Group works in collaboration with the Pain and Symptom Control Task Force specifically to advise on evaluation of patients' subjective experience of symptoms.
3. To advise on the design, implementation and analysis of quality of life studies within EORTC trials, in cooperationwith the Quality of Life Unit at the EORTC Data Center.
4. To conduct basic research in quality of life assessment.
5. To contribute to teaching/training initiatives to promote the EORTC approach to quality of life assessment, e.g. through preparation of teaching material, oral presentations, etc.
6. To develop and maintain liaison with other non-EORTC groups conducting quality of life studies in oncology e.g. NCI-Canada Clinical Trials Group.

2. INTRODUCTION

When evaluating the efficacy of medical treatment on cancer, prolongation of life expectancy and tumor shrinkage have traditionally been taken as outcome measures. Despite the substantial side effects and functional impairment often associated with cancer treatment it is only recently that attention has been given to the assessment of quality of life (QoL). This increasing interest in QoL has important implications for clinical trials, as careful planning is required at all stages of a study from protocol design through to reporting of results. The EORTC Quality of Life Group (QLG) and the EORTC Quality of Life Unit (QLU) want to enhance the quality of data that is collected, and this is best achieved through sharing expertise and cooperating and collaborating with the Clinical Groups. It is important that the relative roles of each of the key players are clearly defined and understood. A common set of guidelines allowing a systematic approach across all EORTC clinical trials should further enhance the quality of data that is collected.

This manual aims to provide guidance for standardizing QoL assessment across EORTC randomized clinical trials. The roles of the EORTC QLG, QLU and Clinical Groups are described, along with the current procedures to be followed and the protocol requirements when preparing an EORTC study which includes the evaluation of QoL. Information is given to help decide when QoL assessment is likely to be a relevant and useful endpoint in a clinical trial, and to select appropriate instruments to measure QoL. There is a discussion on when and how often QoL should be assessed along with some practical methods for enhancing compliance and distributing and retrieving questionnaires. Finally there are chapters on data analysis and ethical considerations.

Whilst the guidelines primarily relate to EORTC phase III trials they may be of interest in phase II trials and to other trialists outside the EORTC.

3. THE PLAYERS

3.1. EORTC Quality Of Life Group

The EORTC Quality of Life Group (QLG) was created in 1980. Members of the group include social scientists, clinicians, statisticians, nurses and data managers from both Europe and Canada. Initially the group's activities centered on promoting the relevance of QoL in clinical trials and advocating its measurement, but progressed in 1993 to the development of a valid and reliable instrument for assessing QoL. A modular approach was adopted with a core questionnaire, the QLQ-C30, supplemented by disease and treatment specific questionnaires or "modules" (Aaronson et al., 1993; Bjordal et al., 2000) (Appendix 1). Considerable emphasis was placed on cross cultural applicability and the current version of the core questionnaire is now available in many languages (Appendix 2).

Permission to use the QLQ-C30 may be obtained from the QLU and there is no charge for academic users. A scoring manual is also available (Fayers et al., 2001). (See Appendix 3 for details on how to contact the QLU.) In collaboration with the QLU the QLG has produced a reference manual (Fayers et al., 1998b) in which datasets from various trials are pooled to provide tables and graphs of QoL scores for groups of patients, stratified by well defined variables such as type and stage of cancer, age, gender.
The manual is available in printed form or on a CD-ROM from the QLU.

All modules are developed according to strict guidelines (Blazeby et al., 2001) and are subject to peer review. If a module has been fully validated and published then permission to use it may be obtained from the QLU. Otherwise investigators should contact the QLU for the name and address of the principal coordinator. For a list of currently available modules see Appendix 4.
Translation guidelines have been produced (Cull et al., 1998) and may be obtained from the QLU. As part of the development process all modules are field tested in English and at least three other European languages.
Recently, a database containing all items from the core questionnaire and all existing modules in the available languages has been developed by the QLU (Vachalec et al., 2001). This Item Bank is accessible through the Internet (www.eortc.be/itembank). A user name and password may be requested from the QLU.

Research activities of the group now fall into four main categories:

1. Further module development.
2. Joint scientific projects with EORTC Clinical Groups where QoL is an endpoint in a new clinical trial protocol.
3. Statistical/Methodological issues.
4. Other areas.

A joint scientific sub-committee has been formed consisting of QLG members and QLU staff, all with expertise in different disease sites in addition to their knowledge of QoL assessment.
The aim of the sub-committee is to be able to offer advice to every EORTC Clinical Group on incorporating measurement of QoL into a clinical trial protocol, and on some of the practical issues associated with implementation. The aim of this manual is to ensure consistency of advice across EORTC Clinical Groups. For a list of joint scientific committee members and their contact address see Appendix 5. Where no suitable contact is listed the QLU will be responsible for all activities.

3.2. EORTC Quality Of Life Unit

The rapid growth in the number of studies assessing QoL emphasizes the need for a coherent policy and a standard approach to conducting this research. It is for this reason that the Quality of Life Unit (QLU) was created in the EORTC Data Center in November 1993. The Unit's main objective is to stimulate, enhance, and coordinate QoL as a treatment outcome in cancer clinical trials. In this context, the principal tasks of the Unit are to establish an adequate infrastructure for the data management of QoL studies; to facilitate the incorporation of QoL data collection into clinical trial protocols, and to support the analysis of QoL data in EORTC clinical trials. Both the QLQ-C30 and the modules are copyrighted instruments developed by the QLG with all rights reserved. Written prior consent of the QLG is therefore required for its use and the administration of the QLQ-C30 is an additional responsibility of the QLU.

Since its creation, the Unit's tasks have expanded, and staff numbers have increased. Staff in the Unit now includes a coordinator, a statistician, a data manager, an administrative assistant and research fellows. The translation coordinator, the module development manager, and the QoL specialist, also working at the QLU perform more specific support tasks.

The QLU is involved in a wide range of studies, across EORTC Clinical Groups, through all phases from the design to the analysis and publication of the results. The Unit has responsibility for supervising the data management for QoL evaluations in EORTC studies, and where possible encourages investigators to adopt a standard approach. This is achieved by involving the QLU in reviewing QoL issues in new protocols before they are submitted to the EORTC Protocol Review Committee. To ensure adequate rates of patient accrual, compliance, and data quality, there is a continuous need to maintain and improve standard data management strategies. Training courses are considered as an important preparation for those responsible for data collection in the clinical setting, and the QLU in collaboration with the QLG actively pursues such activities. A standard list is available for coding information on the reasons for missing data and incomplete forms, and rules have been drawn up for coding missing or ambiguous data (Appendix 6).
In addition the QLU provides Clinical Groups with regular feed-back on compliance figures, which should be prepared by the Clinical Group's data manager or statistician. For a list of all current EORTC studies that include QoL as an outcome measure see Appendix 7.

Statistical research activities at the QLU include collaboration with the QLG on production of the reference values data manual and investigating various methods of analyzing QoL data in cancer clinical trials. Analyzing QoL data may be complicated for several reasons e.g. repeated measures are obtained, data may be collected on ordered categorical response scales, the instrument may have multidimensional scales and complete data may not be available for all patients. In addition, it may be necessary to integrate QoL with length of life. The QLU has made some progress in all of these areas and has published articles in peer reviewed journals in association with members of the QLG and with members of other national cancer research organizations (Curran et al., 1998a; Curran et al., 1998b; Rosendahl et al., 1997; Troxel et al., 1998). The QLU has also developed statistical expertise from analyzing data from various EORTC clinical trials (Curran et al., 1997; Curran et al., 1998c).

3.3. EORTC Clinical Groups

The fundamental structure of the EORTC Treatment Division is based upon 28 Clinical research Groups and four task forces which develop their clinical research through the direct input of their participating scientists. Research is accomplished mainly through the execution of large, prospective, randomized, multinational cancer clinical trials. More than 2,500 clinicians located in 350 medical institutions in 35 countries participate in EORTC protocols. Each year approximately 6,500 new patients are entered into about 100 ongoing studies. Although there are 28 treatment Clinical Groups in the EORTC, not all of them have included QoL as an outcome measure in their trials. However, some groups have a long-standing experience in QoL assessments.

4. PROTOCOL DEVELOPMENT PROCEDURE

4.1. Background

The EORTC New Treatment Committee (NTC) and the EORTC Protocol Review Committee (PRC) are comprised of clinical trial experts including medical doctors, statisticians and QoL researchers. The PRC may also avail itself of external consultants who are specialists in a given field and to whom protocols may be submitted for external review.

The NTC reviews and approves the concept of EORTC trials with non-registered modalities on the basis of their scientific background, interest and feasibility. The PRC performs these same functions for EORTC trials with registered modalities, but additionally reviews the methodology for all studies regardless of modality. It also verifies that important scientific, methodological, collaborative, and administrative issues are in agreement with general EORTC operating procedures. Thus the roles of the NTC and PRC are two-fold:

1. To review and approve all new EORTC protocols with respect to their scientific value, feasibility and relevance within the framework of the EORTC.
2. To assist the Clinical Groups whenever necessary concerning any aspect of the design and implementation of their studies.

4.2. Submission Procedures

Within the EORTC, proposals for conducting a new clinical trial are developed by the Clinical Groups who generally appoint a writing committee to prepare the protocol. One investigator, appointed the Study Coordinator, will actually write the protocol and is responsible for the good conduct of the study within the Clinical Group. The protocol must be written in accordance with EORTC guidelines. Guidelines for submission of outlines and protocols to the EORTC NTC/PRC are provided on the EORTC website at www.eortc.be. For studies which include QoL as an endpoint the protocol writing committee consists of the study coordinator and additional members of the Clinical Group of the particular disease site, the statistician and data manager of the Clinical Group, a liaison person from the QLG and/or staff from the QLU. For all newly proposed phase III trials, a study outline describing the trial must be submitted to the NTC/PRC. The study outline template is available on the EORTC website (www.eortc.be). It is amended periodically. The current outline appears in the format of a standardized questionnaire. One section of the outline is specifically related to QoL as follows:

Do you intend to assess QoL in the study?

Yes/No

If yes,

- Have you contacted the Quality of Life Group (liaison person)?

  Yes / No

- What is the rationale for including quality of life in the study?
- Which QoL instrument(s) will be used in the study?

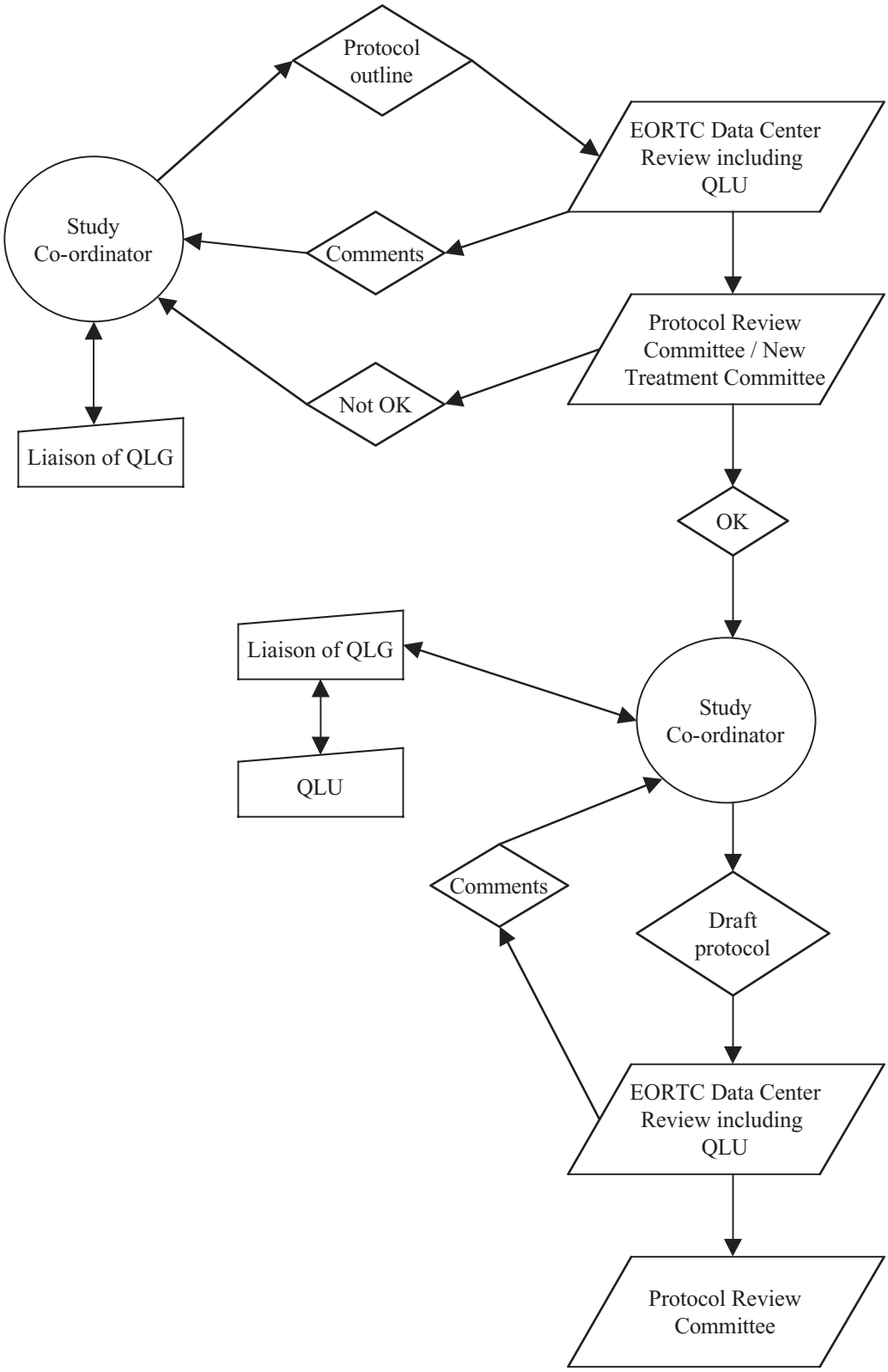
Issues relating to QoL must be discussed with the liaison person from the QLG or the coordinator of the QLU *before submission*. When the outline is complete it may be submitted electronically to the Data Center. Prior to submission to the NTC/PRC the outline is reviewed internally by the appropriate EORTC Data Center personnel, including the statistician, the medical supervisor and the QLU. After internal review the collective comments of the EORTC Data Center personnel are sent to the study coordinator. When the comments of the EORTC Data Center personnel are taken into account the revised outline is sent to the NTC/PRC.

The outline is sent for external review. Afterwards, the study coordinator will receive a letter prepared by the Chairman of the PRC/NTC stating the results of the outline review: accepted, accepted pending modification, to be revised and resubmitted, or rejected. The PRC meets quarterly to discuss “problem projects”. In order to have an outline discussed at the PRC meeting, it should be submitted at least six weeks prior to the meeting.

After approval of the basic concept by the PRC/NTC the study coordinator is encouraged to develop the full protocol in association with all members of the writing committee. The liaison representative of the QLG in conjunction with the QLU will draft the sections covering the topics presented in Table 1. After agreement from all parties involved (the liaison person of the QLG, the QLU and the study coordinator) on the content of the sections of the protocol related to QoL assessment, the text is incorporated into the protocol. Prior to submission of the final protocol to the PRC, the version of the protocol to be submitted to the PRC must be approved by the EORTC Data Center.

For phase II protocols written according to a PRC approved master protocol, a "quick procedure" may be employed. This implies that the first step involving developing the two-page outline is bypassed, and the full protocol may be developed in conjunction with the Data Center and the liaison person of the QLG.

Figure 1: Flow Chart Showing Stages in Protocol Development Prior to PRC Submission



5. PROTOCOL REQUIREMENTS

The success or failure of a trial may depend on how well the protocol was designed and written. The protocol must be detailed and precisely worded with all the requirements clearly indicated so that the study may be uniformly carried out by all participants (Fayers et al., 1997). Table 1 lists the topics that should be covered in the full EORTC protocol. Discussion of these topics can be found in the relevant chapters.

Table 1: Protocol Contents Relevant to QoL Assessment

	Topic	Chapter
1.	Description of rationale for measuring quality of life	5
2.	Statement of quality of life variables considered relevant (which side effects, late effects, psychological domains, primary or secondary endpoint)	6
3.	Detailed description of design of the study	7
4.	Patient eligibility	
5.	Choice of instrument (which and why)	6
6.	Timing of assessments	7
7.	Mode of data collection (in person, by mail, etc.)	9
8.	Statistical considerations (sample size, hypothesis to test)	10
9.	Missing data (importance and methods for enhancing compliance)	8, 9, 10
10.	Informed consent procedure	11
11.	Appendices (instruments, patient information leaflets, consent form, diary record system)	

6. SELECTING TRIALS IN WHICH QUALITY OF LIFE IS RELEVANT

It would be unrealistic to recommend that QoL should be evaluated in all clinical trials. QoL has considerable resource implications and these should be balanced against the usefulness of the data and its likely impact on recommendations for treatment choice following completion of the trial.

6.1. Phase I & Phase II trials

Most investigators acknowledge that QoL measurement is unnecessary in Phase I and Phase II trials because the primary aim of such studies is to determine anti-cancer activity and toxicity, and patient numbers are usually small. However, it may still be useful in the following circumstances:

- If a QoL instrument needs piloting before being used in a Phase III trial. Investigators can ensure that the instrument covers all the relevant issues, assess reliability and validity and test the infrastructure for future data collection.
- As an exploratory study to investigate if there are unexpected QoL issues not covered by the questionnaire in use. Interventions may then be required in a phase III study to minimize symptoms and dysfunction.
- If a randomized study is likely to continue as a phase III study in which QoL is considered an important outcome measure.

6.2. Phase III trials

Some advocate that QoL should be measured in all phase III cancer trials (Osoba, 1992), and investigators are required to justify a decision not to assess QoL. Others adopt a more pragmatic approach and select trials where QoL is particularly relevant (Aaronson, 1995). Gotay and colleagues (Gotay et al., 1992) recommend that QoL evaluation is included in the following settings:

- A trial where QoL is considered to be the primary endpoint (e.g. the comparison of two palliative treatments).
- A trial where no significant differences between treatments are expected in terms of cure, disease free survival or overall survival but one arm is expected to be associated with significantly more morbidity. Following the trial the decision as to which treatment to recommend may have to be based on QoL outcomes.
- A trial where survival and disease free survival or cure are expected to differ between the two arms but the advantageous primary outcome is only achieved at the expense of major toxicity, e.g. high dose chemotherapy plus bone marrow transplant versus standard chemotherapy. Here data on QoL assessment can be used to support decision making when the benefits identified in the primary endpoint have to be weighed against a negative outcome in terms of QoL.
- It may also be necessary to assess QoL in studies evaluating cost-effectiveness. Specialist measures or instruments will usually be required and the advice of a health economist through the Health Economics Unit at the EORTC Data Center is recommended.

7. SELECTION OF INSTRUMENTS

Any questionnaire chosen to evaluate QoL should have proven, good psychometric properties with respect to validity, reliability, and responsiveness to change (Ware, 1987). Responsiveness refers to a combination of both reliability (identical scores in stable subjects over time) and sensitivity (the ability to demonstrate changes when the subject's state of health improves or deteriorates, or to detect treatment effects). This latter characteristic is particularly important in a clinical trial setting. The questionnaire should also be simple, brief, and easy to administer. These properties enhance participation and compliance,

and they reduce the burden for both patient and staff. Gelber & Gelber (Gelber & Gelber, 1995) recommend that the selection of instruments should be based on an assessment of the following four issues:

1. The purpose of the clinical trial.
2. The patient population.
3. The treatments and their potential toxicities.
4. The resources of the investigators and the participating clinicians.

In addition the QoL questionnaire should be available in the appropriate languages in relation to potential participants in the clinical trial.

There are two basic types of instruments: generic and disease specific. Generic instruments focus on the main components that constitute QoL, and they are intended to be applied in a wide range of populations and health states across all diseases. Disease specific instruments have been developed especially to detect subtle, disease and/or treatment related effects. There are many excellent validated self-completion questionnaires for cancer patients available e.g. EORTC QoL Questionnaire (EORTC QLQ-C30), Functional Assessment of Cancer Therapy (FACT), Rotterdam Symptom Checklist (RSCL), Functional Living Index-Cancer (FLIC) (Aaronson, 1995; Cella et al., 1993; de Haes et al., 1990; Schipper et al., 1984). All these questionnaires are multidimensional, minimally covering physical, psychological and social domains as well as some overall judgement of the valuation of life or the health condition. It is rarely necessary (or advisable) to develop a new instrument.

7.1. EORTC QLQ-C30

For trials coordinated by the EORTC both the QLG and the QLU recommend that, whenever possible, QLQ-C30 (version 3) should be used in its entirety for a number of reasons:

- The instrument has been carefully developed in a multi-cultural setting.
- Translations are available in 43 languages. If additional translations are required they can be developed using rigorous and standardized translation procedures.
- The instrument has been shown to be valid, reliable and responsive to change.
- Disease specific modules are available to supplement the core questionnaire.
- Study results can be compared across trials.
- Reference data is available for calculating sample sizes.
- The questionnaire is easily understood by most patients and is quick to complete (mean time 11 minutes).

Table 2: Structure of the EORTC QLQ-C30

EORTC QLQ-C30 (30 Questions in total)			
FUNCTIONAL SCALES (16 questions)	SYMPTOM SCALES (6 questions)	SINGLE ITEMS (6 questions)	GLOBAL QUALITY OF LIFE (2 questions)
Physical	Fatigue	Constipation	Global QoL
Role	Pain	Diarrhoea	
Cognitive	Nausea/vomiting	Sleep	
Emotional		Dyspnoea	
Social	Appetite		
		Financial	

Modification of the questionnaire is not permitted without prior consent of the QLG as it is a breach of copyright. Scales that appear irrelevant should only be omitted in exceptional circumstances and individual questions should never be used alone when they form part of a scale. Amongst the reasons for this are:

- The psychometric performances of individual scales and items when used alone are not known.
- Less than 5% of patients find any one item upsetting.
- Having been used in over 600 studies there is no evidence to suggest that patients are bothered by questions relating to symptoms or problems they do not experience.
- Unexpected and important results may be observed and used to generate hypotheses for future studies.
- Data can be used for updating the reference manual of normative data.
- Data can be used for meta analysis.

In collaborative studies with other groups an alternative measure may be proposed. This may occur when industry has sponsored a particular instrument or the other party has had experience with one. In such cases it is recommended that, finance permitting, a head to head comparison with the other instrument is carried out (e.g. FACT, RSCL, SF36).

7.2. Modules

Cancer site or treatment specific modules have been developed by, or in collaboration with, members of the QLG according to strict guidelines drawn up by group members. These modules are in various stages of development and a list is included in Appendix 4 along with contact details for the principal developer. For modules where validation data has been published, permission to use the module may be obtained from the QLU, otherwise permission may be obtained from the principal developer. The modules are intended to supplement the QLQ-C30 and should not normally be used without concurrently administering the QLQ-C30. Where an EORTC module is not available, use of an existing instrument in the area of interest, with known psychometric properties, is preferred to developing an ad hoc questionnaire.

Where a specific research question is posed and use of the QLQ-C30 and a module is insufficient, it is possible to add extra questions as a checklist. The QLU has recently set up the Item Bank in close collaboration with the QLG Item Bank Committee. Validated items from the Item Bank can be used

to supplement the core questionnaire and other modules, if necessary. The use of additional items has to be carefully discussed with and approved by the QLU. Other main aims of the Item Bank are to improve the quality of modules and to standardize the wording of items in existing and future modules, to improve the speed and quality of module development and to improve the speed of translations.

7.3. Other Situations

Whilst the QLQ-C30 is usually the instrument of choice in phase III clinical trials, there are situations where its sole use would be inadequate or inappropriate as it does not address all the research questions of interest. e.g.

- Survivorship studies where the QLQ-C30 does not cover the relevant aspects such as work rehabilitation, relationships, infertility.
- Cost-effectiveness analyses where a single measurement is often required.
- Paediatric studies.

8. WHEN & HOW OFTEN SHOULD QUALITY OF LIFE BE ASSESSED?

Within the context of a clinical trial where it has been decided that it is appropriate to assess QoL there are countless opportunities when patients could be asked to complete a QoL measure. However the burden on the patient and the associated data management and data analysis problems necessitate some limitation. The ideal number and timing will vary from one clinical trial to the next, dependent upon the research hypothesis, but will usually involve assessments before, during and after treatment. To facilitate a sensible and practical interpretation of the results the minimum number of measurements should be used. They should be timed to yield maximum information about changes in QoL due to both treatment and changing disease status. Timing schedules should be similar across all treatment arms.

8.1. Before Treatment

Information on the patient's QoL prior to their diagnosis of cancer is rarely available so a pre-treatment score is usually considered to be their starting point or baseline assessment.

- It allows for comparison between study groups before treatment is initiated. If differences are found they can sometimes be controlled for during subsequent analysis.
- A pre-randomization assessment provides a starting point for assessing changes caused by both treatment and disease status.
- Where follow up data is missing and the patient was known to still be alive it may allow for detection of a systematic bias. For example, patients with a poor QoL at baseline may not always be asked to complete follow up assessments.
- In addition it has also been shown that QoL at baseline may be of use as a prognostic factor for clinical outcomes, including survival, response to treatment and nausea and vomiting (Coates et al., 1997; Osoba et al., 1994; Tannock et al., 1996).

8.2. During Treatment

Choosing a schedule for collecting QoL data during treatment will often involve a compromise. To reduce the administrative burden and thus improve compliance, assessment times should coincide with the clinical care schedule dictated by the trial regimens. However assessments should be timed to reflect the expected profile of treatment burden and toxicity. These requirements often conflict and it is often not possible to recommend a fixed schedule to be used in all trials. Rather it is essential to appreciate the clinical course of the disease and to discuss within the protocol writing committee expectations regarding serious or acute

effects, stable periods and chronic problems. An appropriate research hypothesis can then be formulated and used as a guide to determine appropriate measurement times.
When writing a protocol one should aim for similar (if not identical) schedules in terms of frequency and number of assessments between the two arms.

8.2.1 Anchoring Events

Assessments can be a) time-based - given a set number of days or weeks after randomization, independently of the specific treatment schedules, or b) event based - given to coincide with specific treatment cycles (See figures 2 and 3), or even daily. A decision as to which is the most appropriate approach will depend upon the research question, but careful thought should be given to the consequences of delayed treatments.

- An event-based approach is often logistically easier to manage but it does not provide QoL data between treatments. Assessments are usually scheduled to take place immediately before the next course of treatment. In the case of treatment delays assessments would still fall immediately before a specified course but if the reason for delay was toxicity, information will not be collected on the patient's QoL at the time of the delay. If the "events" happen at different times in the trial arms but the treatments are equivalent, patients in one arm may have more advanced disease because a longer interval has elapsed since randomization (e.g. figure 2, 3rd QoL assessment). The number of assessments is independent of the duration of treatment, which may simplify analysis.
- Whilst it may be more difficult to facilitate reliable data collection in a time-based approach, if the times are carefully chosen data on QoL between treatments can be made available. Assessments will always be at the same time interval in all arms relative to randomization. The number of assessments is only known if treatment is not delayed and analysis of data may become more complex.
- Sometimes a combination of the two is appropriate. For instance recommending that QoL should be assessed at day 21 of each cycle before the next treatment is administered ensures only one assessment per course, but should be sensitive to changes in QoL at a time of treatment delay if this is necessary.
- Daily assessments may uncover details that would be missed by less regular assessments, but there may be difficulties with compliance. A large volume of data is generated and a corresponding increase in time must be allowed for input and analysis. The use of daily or weekly measurements should be very limited.

8.2.2 Time Scale

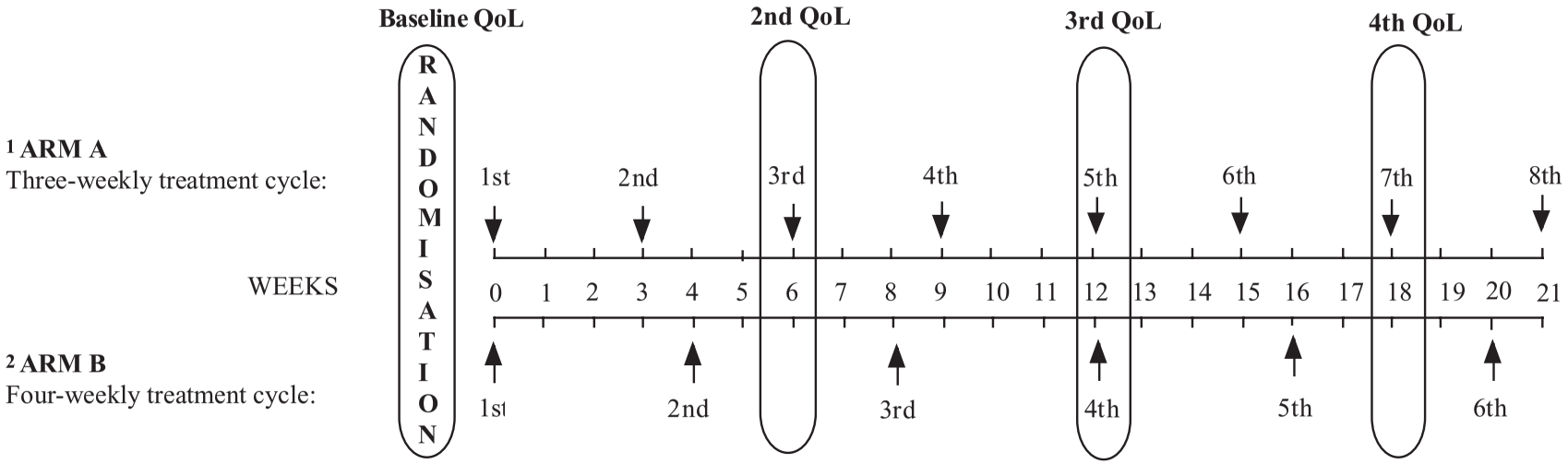
The time scale of the chosen questionnaire needs to be considered. Some questionnaires refer to symptoms and QoL during the previous week (e.g. EORTC QLQ-C30) whilst others ask about current status. Where acute side effects are expected within a few days of treatment it may be inappropriate to collect QoL data three weeks later on the patient's next visit.

8.3. After Treatment

Once treatment is completed the number of QoL assessments required and their frequency depends upon the research hypothesis and whether QoL was specified as the primary endpoint. As before, a compromise will be needed to balance "exploratory" requirements with more pragmatic considerations. To eliminate bias, assessments should occur at equal times in each arm relative to randomization and not to end of treatment.

- In clinical trials where the patients have a poor prognosis data may be lost if the interval between assessments is too long. Care should be taken not to overburden patients in the last few months of their lives.

Figure 2: "Time" Based quality of life evaluation
e.g. 6 weekly assessment

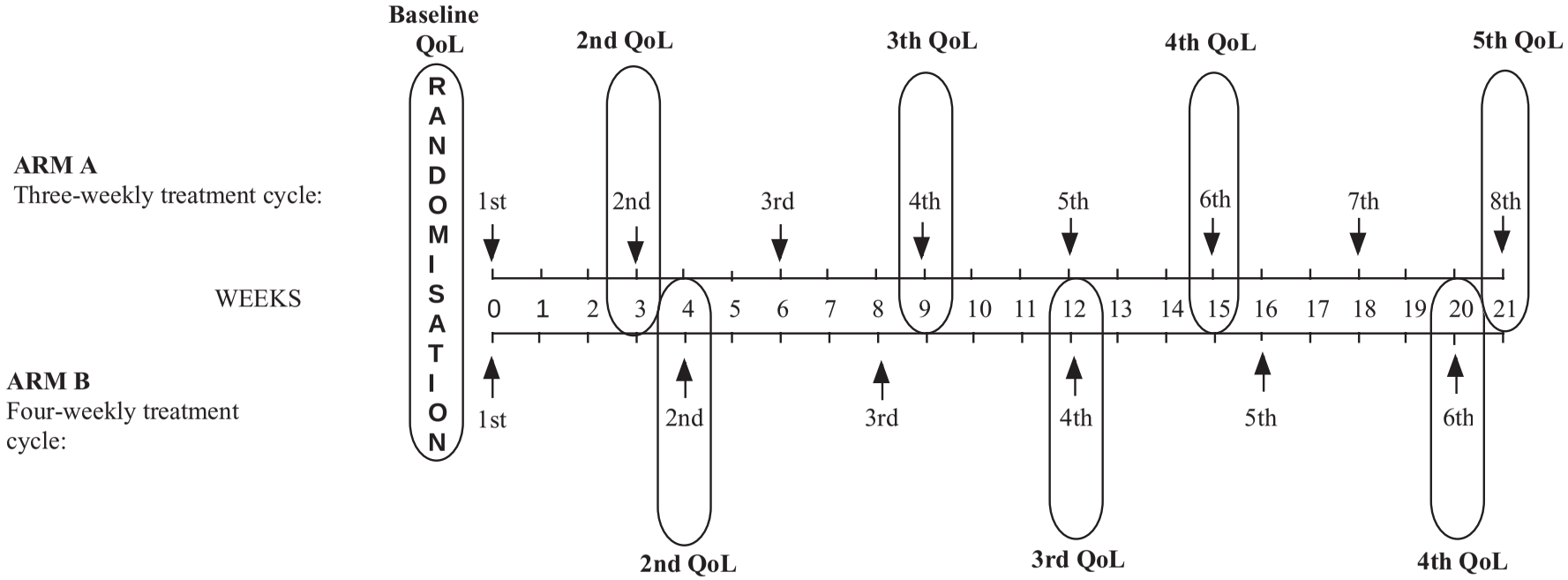


In both arms the baseline assessment is carried out before randomisation. Treatment should start as soon as possible after randomisation. Thereafter assessments are at 6 weekly intervals relative to the randomisation date.

1 In Arm A treatment is three-weekly so QoL assessment always coincides with the start of a new cycle.

2 In Arm B treatment is four-weekly so QoL assessment sometimes falls between cycles and sometimes it coincides with the start of a new cycle.

Figure 3: "Event" Based quality of life evaluation
e.g. Assessment every 2nd cycle



QoL is assessed at baseline (before randomisation) in both arms. Treatment should start as soon as possible after randomisation and QoL is then assessed after the 2nd, 4th, 6th (and 8th) cycles.

In both arms QoL assessment always coincides with the start of a new cycle but in arm A there are more assessments and assessment is also more frequent.

- For patients receiving radical treatment where long-term survival benefits are expected intervals between assessments can be extended.
- If QoL data is required at the time of relapse local institutions should consider how they will achieve this. The patient will be withdrawn from the study at an unknown point in time and the appropriate questionnaire may not be to hand, but compliant patients may be upset if, as they relapse, an assessment of their QoL is seen as unimportant. Once the patient has relapsed subsequent QoL data collection may be hampered by the infrequent, if any, visits, of patients to the study hospital.
- In some studies it may be necessary to collect QoL data until death. For example studies comparing immediate versus delayed treatment, or studies when a treatment is expected to prolong time-to-progression or disease-free interval but no difference in duration of survival is expected. In these situations it is important to illustrate the benefits of extending the time to an event versus the side-effects of the treatment.
- If questionnaires are to be handed out in person then the assessment schedule needs to coincide with the follow-up schedule. However if questionnaires are to be mailed any appropriate time schedule can be used.

9. ENHANCING COMPLIANCE

Unless or until collecting QoL data is seen as part of routine clinical practice it will be necessary to implement specific measures for each clinical trial because QoL data cannot be collected retrospectively. Missing data makes analysis very complicated and results difficult to interpret. If QoL is specified as an important primary or secondary endpoint in a multicenter clinical trial protocol then it must be mandatory in all participating centers. If QoL assessment is left as optional and restricted to centers that have an infrastructure to facilitate it, the patients may not be representative of the wider sample drawn from all participating centers. QoL assessment in a subgroup of patients would still require a large number of patients in order to satisfy the statistical power of the study. There would be a crucial need for maximal compliance in the subgroup and an unforeseen number of drop-outs might render the study unevaluable.

Both staff and patients should be provided with the necessary resources for optimal data collection at the appropriate times. Collaboration between the study coordinator, the data center responsible for the day to day administration of the study and the individual investigators can lead to organizational improvements. More specific measures can be targeted at the patient, the physician and the data manager or the research nurse, though there is considerable overlap.

9.1. Organizational Issues

Good organization and forward planning ensures that all those involved know their respective roles.

- A small local pilot study may be organized prior to commencement of the main study, followed by a debriefing meeting. This will enable each center to make a realistic assessment of the number of patients they are likely to recruit and the time, space, personnel and financial resources that will be required.
- Where the task of collecting QoL data is shared amongst a number of people, one should be appointed as the local coordinator. The coordinating data center could request details of this individual at the same time as they verify ethical committee approval and collect data on laboratory normal values. The individual is then responsible to the coordinating data center and any queries can be directed through them.
- Baseline QoL assessment should be one of the eligibility criteria. Completion of a QoL questionnaire would then be included on the checklist which has to be completed before randomization can take place.

- The procedures for collecting QoL data at each center should be documented and include names and contacts for all those involved. Then in the event of staff absences everyone is aware of where to find the relevant paperwork and who has what responsibility.
- Recruitment and compliance figures are prepared by the data manager and statistician of the Clinical Group and can be presented by the QLG liaison person or a representative of the QLU every six months at the Clinical Group meetings. This allows feedback and discussion if there are any problems. Procedures in place within the EORTC to monitor data timeliness apply to QoL data as well as clinical data. Investigators with patients who are not evaluable due to missing QoL data will be notified so that policies can be set in place within their institution to rectify the problems and not jeopardize the study.
- For some trials the QLU has developed a schedule for QoL assessment for each patient starting from the date of randomization. This is sent to the attending physician and kept in the patient files. This is useful for the medical staff to check when the patient should complete the QoL assessment. It also allows the study monitor to check if QoL assessments have been done according to the schedule provided in the protocol.
- A spare copy of the QoL questionnaire should be kept in the patient's clinical file; if the questionnaire is lost a backup copy will then be available.
- The following set of questions are included on the clinical Case Report Forms as a reminder that QoL questionnaires should be completed and to aid in determining the reasons for missing questionnaires:

Has the patient filled in the current QoL questionnaires, 0 = no, 1 = yes
If no, please state the main reason

1 = patient felt too ill
2 = clinician or nurse felt the patient was too ill
3 = patient felt it was inconvenient, takes too much time
4 = patient felt it was a violation of privacy
5 = patient didn't understand the actual language / illiterate
6 = administrative failure to distribute the questionnaire to the patient
7 = not required at this time point
8 = other, please specify

9.2. The Patient

Most patients are willing to complete QoL questionnaires. Specific measures to improve compliance include:

- Providing a clear explanation of the reason for collecting QoL data in the context of the rest of the study. This information should be given verbally and supported by a written information sheet.
- Providing information on when questionnaires will be due e.g. a copy of the reporting schedule mentioned above.
- Informing patients what will happen to their completed questionnaires, e.g. they will not be stored in the patient's clinical notes and will remain confidential. (Within the field of clinical trials there are major discussions about individual and collective ethics. In many studies completed QoL questionnaires from trial patients are not made available to the treating clinician during their consultations and patients should therefore be made aware of this at the time they consent.)
- Ensuring the questionnaire itself is not too long and contains questions that appear relevant to the patient and are easily understandable. If more than one questionnaire is used care should be taken to avoid duplication of issues. The format and layout should be clear and include written instructions.
- Providing a private and comfortable environment for completing the questionnaire.

- Providing help if necessary for patients who are unable to complete the questionnaire unaided for whatever reason (e.g. poor comprehension, no glasses, etc.).
- Showing appreciation once the questionnaire is completed and expressing an interest in any concerns the patient may raise.

9.3. The Physician

Some clinicians are unconvinced of the scientific validity of QoL assessment. They may be sceptical about the value of measuring QoL within a clinical trial and therefore have difficulty in seeking the cooperation of all patients. The following measures may have a positive influence on their opinion of QoL assessment:

- QoL data collection should not be presented as an "optional extra" in a trial but rather seen as a mandatory and integral part of the study.
- Published work where QoL data has made a significant contribution to the scientific validity of a study should be presented and promoted.
- The study coordinators should be seen to be convinced of the value of QoL measurement. They should also be clear as to the rationale for collecting QoL data in their particular trial and should use this information to motivate the study investigators.
- Clinical considerations suggested during discussions between the coordinator and the investigators, should, where possible, be taken into account when designing the QoL component of the trial.
- Investigators should receive feedback regarding the QoL data collection in much the same way as they receive recruitment updates and preliminary results from the clinical component of the study.

9.4. The Data Manager/Nurse

Responsibility for distributing QoL questionnaires is often allocated to a data manager or research nurse. A distinction should be made between someone who is available in the clinic to attend to the patient personally, and someone who visits the treatment center at regular intervals, but can only leave the questionnaire in a prominent place with a reminder that it be given to the patient on their next appropriate visit and collected later. The latter have a limited role to play.

- If data managers or nurses are expected to distribute questionnaires personally they should be well informed so that they can answer questions. Trial specific workshops held prior to the commencement of a study have been advocated. The rationale for collecting QoL data can be explained and the practical procedures to be followed discussed in detail. In a multi-national setting this may not be practical or cost-effective. As an alternative, national training courses in data management could be encouraged to broaden their coverage of general issues surrounding the collection of QoL data, which would then be applicable to a wide variety of trials.
- Regular contact between the data manager/nurse and the study investigator should increase motivation and enhance compliance.
- When the investigator receives feedback on the center's compliance they can convey this information to the data manager/nurse.

For some patients, completing a QoL questionnaire may prompt them to seek more information or support. It is then important that the data manager/nurse is competent to deal with any issues that may arise, or knows where to refer the patient for appropriate help. Relevant training should be arranged along with an awareness of the resources available.

10. PRACTICAL PROCEDURES FOR DATA COLLECTION

The EORTC QLU have produced a two page information sheet "EORTC Guidelines for administration of QoL questionnaires". (Appendix 8)

10.1. Mode Of Delivery

The two most common modes of administration are handing questionnaires to the patient in person whilst they attend a clinic or mailing them to their home address. For baseline and on-treatment assessments the QLG and QLU recommend handing out questionnaires in person because:

- The first time patients are asked to complete a questionnaire they may not understand the instructions or may find some questions confusing. Available staff can give a verbal explanation.
- Where patients are unable to fill in the questionnaire themselves for practical reasons (e.g. forgot glasses, too frail) the member of staff may choose to read out the questions and fill in the questionnaire on the patient's behalf. This should then be recorded on the form.
- There is an opportunity to check questionnaires for missing data and ascertain whether this is accidental or deliberate. In the former case patients can be asked to complete the missing questions whilst in the latter case the questionnaire should be marked that the patient did not wish to answer particular questions.
- Some patients are unable or refuse to complete a whole questionnaire; the reason for this can then be ascertained and recorded.
- When necessary one can tactfully try to discourage relatives from answering the questions on behalf of the patient.

Patients should be discouraged from taking questionnaires home and returning them either by post or in person on their next visit. (In such cases the investigator has no control over the exact completion date or whether the patient's answers were influenced by family members or friends.)

If questionnaires are mailed to patients it will be necessary to check on their survival status beforehand to avoid distressing relatives of patients who have died. Reply paid envelopes should be provided.

10.2. Time Of Delivery

- It is normally recommended that baseline data is collected before randomization, so that completion can then be made an eligibility criterion and the outcome of randomization cannot influence any of the domains in the QoL score.
- As it is preferable to reduce all sources of potential bias it is recommended that questionnaires are completed prior to seeing the physician. This has the advantage that it may prompt the patient to discuss any worrying symptoms.

10.3. Missing Data

Protocols should contain explicit instructions for normal practice and what to do in the event of a protocol deviation. If a questionnaire is missed the protocol should be clear on the practice to follow. Should the data be accepted as missing and only a reason recorded (e.g refusal, nurse forgot, etc.) or should the patient be contacted? Options include mailing the questionnaire and a reply paid envelope or telephoning. Whichever method is chosen it is important to establish beforehand an acceptable time delay or "window" during which the questionnaire must be completed. During treatment, windows should be narrow to evaluate short term toxicity timing (e.g. +/- 1 week) but during follow up wider windows may be acceptable.

10.4. Proxy Ratings

It is generally agreed that patients are the best raters of their own QoL (Slevin et al., 1990). There are circumstances in which it is difficult or even impossible for patients to rate their own QoL (e.g. patients who are cognitively impaired due to their cancer, patients who are terminally ill and children). In these circumstances their QoL may be assessed by a third person. This can be a family member (e.g. partner or a parent) or the care taker (e.g. physician or nurse).

In two samples of cancer patients, Sneeuw (Sneeuw et al., 1997,1998) examined the level and pattern of agreement between ratings provided by patients and their significant others on the EORTC QLQ-C30. At the individual patient level, more than 90% of scores were within one response category of difference, and correlations for the several dimensions were moderate to good (between 0.40 and 0.80). At the group level, significant others tended to rate the patients as having a lower quality of life than the patients themselves, but this bias was of a limited magnitude.

Sneeuw (Sneeuw et al., 1997) reported very similar findings when comparing ratings provided by cancer patients, significant others and physicians on the COOP/WONCA charts, assessing several quality of life dimensions at a generic level by means of seven single questions. Lower levels of agreement were noted for more private domains, such as feelings, social function, and overall quality of life. The level of agreement between patients and their physicians was only slightly lower than that observed between patients and their significant others. Physicians tended to underrate patients' pain severity.

Stephens (Stephens et al., 1997) investigated the concordance between ratings provided by lung cancer patients and their physicians on eleven symptoms derived from the Rotterdam Symptom Checklist. Of all comparisons made, 78% showed exact agreement between doctor and patient, 18% disagreement by one, 4% by two, and 1% by three grades. However, there was increasing disagreement with increasing symptom severity, and a consistent bias towards doctors underestimating symptom severity. Importantly, physician compliance was higher than patient compliance, and the between-treatment comparisons reached the same conclusions regardless of whether the data was patient-based or physician-based.

If it is anticipated that an increasing percentage of the patient population under study will be unable to complete questionnaires during the course of the trial (e.g. due to neuro-psychological deficits or a seriously deteriorating physical condition) proxy respondents could be considered from the trial outset.

11. DATA ANALYSIS

Analysis of QoL data raises a number of contentious issues, and we outline some of the main ones. Three points are of particular note.

1. The analysis of completed trials will be simpler and more convincing if the principal hypotheses have been specified a priori. Both the hypotheses and the QoL outcomes to which they relate should be specified in detail in the protocol.
2. The definition of "clinically important differences" should be considered at the time of writing the protocol, and should be specified. (This will also be necessary when sample size estimation is based upon QoL endpoints.)
3. Since missing data (non-returned QoL forms) raises major questions about bias and poses severe analytical problems, every attempt should be made to ensure high compliance.

11.1. Simple Comparisons

Many of the complications in analysis arise because studies which assess QoL usually assess each patient at multiple time points. When cross-sectional analyses are carried out (for example, all patients at the pre-randomization time point), many of the problems disappear. Sometimes simple t-tests may be appropriate (for example, when comparing global health status across two treatment arms).

Often non-parametric tests, such as Wilcoxon or Mann-Whitney tests, may be more appropriate because many of the single items and some of the scales have asymmetric distributions. It should also be noted that the single items are mostly 4-point scales, and so ordered logistic regression may be appropriate if one wants to use regression techniques to examine the effect of prognostic variables upon QoL outcomes.

An alternative approach, which may be especially suitable for single items, is to consider percentages rather than means or averages. For example, instead of estimating the average vomiting score for each group of patients, one can calculate the percentage of patients in each group who report "quite a bit" or "very much" vomiting. Many readers may find percentages more intuitive and easier to understand than average levels. For example, the statement “24% of patients reported vomiting at least “quite a bit” has a more obvious interpretation than reports such as “the average level of vomiting was 58.2”. When percentages are used, the analyses often reduce to comparisoıns of binomial proportions or possibly chi-squared tests.

11.2. Multiplicity Of Outcomes

The core QLQ-C30 contains 30 items and a number of scales (five functioning scales, one global health status, and three symptom scales), with the supplementary modules containing additional items and scales. Thus there are potentially many pairwise statistical comparisons that might be made. As is well known, for every 100 independent statistical tests that are carried out, even if we assume there is no treatment effect, we would expect approximately five comparisons to be statistically significant at $p<0.05$. Therefore, when making multiple significance tests, we are likely to obtain about 5% of results as false positives.

There are three main methods of making allowance for this. First and foremost, the study protocol should identify one or two QoL outcomes as being of principal interest. These few outcomes will be the main focus of the analysis, and therefore there will be no problems of multiple testing. It is important that these outcomes are listed in the protocol, to avoid it being suggested that the investigators “cheated” and inspected the data before determining which variables are of most interest. All other analyses may then be regarded as primarily hypothesis generating, and will be regarded more critically.

The second method, which is sometimes used in conjunction with the first, is to adopt "conservative p-values." If many statistical tests are being performed, it is possible to use $p<0.01$ as indicating statistical significance, thereby reducing the rate of false positives. In extreme cases, $p<0.001$ could be used. Rather related to this, "Bonferroni corrections" are often used. The principle underlying this is that in theory one should not use a fixed but arbitrary $p<0.01$ irrespective of the total number of statistical tests. Instead, if it is planned to make N statistical tests, one can estimate the equivalent p-value that will maintain overall significance at, say, 5%. For an overall p-value of a, the Bonferroni method indicates that one should use p-values of a/N for the individual tests (Bland & Altman, 1995). The third method is either to use some form of global multivariate test, or alternatively to reduce the items to a few summary scores. This method has not often been used in QoL studies. Tandon describes applications of global statistics in analyzing QoL data (Tandon, 1990).

11.3. Repeated Measurements

There are various methods available for data description and statistical significance testing when repeated measurements are available for the comparison of two or more treatments. One of the simplest approaches is to use graphical displays and accompany these by cross-sectional analyses at a few specific time points. Ideally, the study protocol will have pre-specified that the analysis will focus upon QoL at these particular time points, with the additional measurements being regarded as of secondary importance. For example, a chemotherapy protocol might specify that differences in QoL at the time of the third course, and also at one month after completion of chemotherapy will be tested for statistical significance. These tests could be accompanied by graphical displays showing the average levels of QoL for the treatment arms and possibly for various patient subgroups.

A second approach is to condense the repeated measurements for each individual into a few summary statistics. For example, one could estimate the average level of each QoL scale, taken over the on-treatment period. This would reduce the repeated on-treatment measurements for each patient to a single score. Other summary statistics that are frequently employed include (a) the overall average QoL for each patient, (b) average QoL after completion of therapy, (c) the worst QoL experienced during therapy (or highest levels of toxicity) and (d) the "area under the curve" (AUC), which is equivalent to the average if the time points are at equal intervals. The analyses can then compare and test the summary statistics. The application of these methods is described by Matthews et al (Matthews et al., 1990).

Finally, some sophisticated statistical methods are available for the analysis of repeated measurement data. Mostly, these methods involve fitting a mathematical model to the data. Since repeated measurements on any one individual are likely to be correlated, the model must allow for the auto-correlation between values at successive time points. The main methods are multivariate analysis of variance (MANOVA) for repeated measures, hierarchical models (multilevel models) and generalized estimating equations (GEE) (Diggle et al., 1994; Goldstein, 1995; Hand & Crowder, 1996; Lindsey, 1993).

11.4. Missing Data

Two types of missing data may be distinguished. First, patients may fail to compete all items on a form, possibly accidentally. The EORTC QLQ-C30 Scoring Manual describes an elementary method of calculating scale-scores when there are a few missing values for some items.

The second, and usually far more serious problem, arises when whole forms are missing. In particular, it is often difficult to know whether patients do not return forms because they feel too ill, or whether the reason is that they feel fine and see little point in replying. Thus one can never be confident that the observed QoL data is representative of all the patients in the study. Sometimes there may be serious bias. Missing data is often a particular problem when carrying out longitudinal (repeated measurements) analysis. However, it should be emphasized that whenever there are many patients with missing data the results of any analysis, cross-sectional or longitudinal, may be suspect. How can we be sure that those patients with data are truly representative of the total sample recruited to the study? Hence, are the results biased?

When data is missing, there is no easy solution for eliminating bias. Therefore, emphasis must always be placed upon avoiding the problems by ensuring optimal compliance with assessment. This cannot be stated too strongly. Any form of correction to the analysis will always be regarded with suspicion by readers, and the study results will only be convincing if compliance is high and missing data is kept to a minimum.

Analytical methods tend to be complex, and are controversial. A special issue of Statistics in Medicine (1998, Volume 17) is devoted to this topic, and contains contributions made on behalf of the EORTC QoL Study Group (Curran et al., 1998b; Fayers et al., 1998a). More recently, Fayers and Machin also published a book on assessment, analysis and interpretation of quality of life data (Fayers and Machin, 2000).

11.5. Interpretation & Clinical Significance

It is relatively easy to obtain a feeling for percentages (for example, "30% of patients reported quite a bit of problem with tiredness"), but many of the items on the QLQ-C30 contribute to multi-item scales which are scored from 0 to 100. Most users are unfamiliar with these particular scales, and do not know how to interpret the mean scores. Also, in a two-arm clinical trial, what interpretation should be given to, for example, a difference between emotional functioning of 58 in one treatment group and 66 in the other? Statistical significance tells us whether the observed data can be explained by chance fluctuations (such as selection of patients), but says nothing about clinical significance. Is a difference of 8 (i.e. 66 - 58) large enough to be important? If a patient's score changes by 8 points, would they even notice the change? Osoba et al. (Osoba et al., 1998) asked patients to complete the QLQ-C30 on repeated occasions, and the patients also rated their perception of change since the previous time they completed the QLQ-C30. Physical functioning, emotional functioning, social functioning and global QoL scales were evaluated. It was found that when these scale scores changed by 5 to 10 points (on the 0-100 scale), patients described

their condition as "a little" better (or worse). A change of 10 to 20 was described as a "moderate" change. A change greater than 20 was "very much" better (or worse).

King (King, 1996) used a very different approach, based upon "known groups" who were expected to differ in terms of QoL scores, such as limited disease patients and those with advanced disease. Data was collated from fourteen published studies. She concluded that for most scales a difference of 5 or less is a "small" difference, but the definition of a "large" difference varied for each scale: for example, it was 16 for global QoL, 27 for physical functioning, and 7 for emotional functioning.

Hjermstad et al. (Hjermstad et al., 1998) report normative data for the QLQ-C30 in a randomly selected sample of 3000 people from the Norwegian population, aged between 18 and 93 years. Data was available for 1965 individuals. Results are presented for the functioning scales, the global QoL scale and the single items. The results are tabulated by age and sex. These normative data may serve as a guideline when interpreting QoL in groups of cancer patients.

For individual patient sub-groups, the EORTC QLG has produced a manual of reference data (Fayers et al., 1998b). Members of the QLG contributed data from their studies, which was pooled for the tables. The manual tabulates the values for QLQ-C30 and its scales according to the main cancer sites divided by stage of disease (early or limited, versus advanced or extensive). Age and gender-specific values are given. This enables investigators to contrast their results with those that have been found in comparable groups of patients.

In summary, the interpretation of results remains essentially qualitative. Clinical significance is subjective, and is a matter of opinion. The values and opinions of individual patients will differ, as will the opinions of the treating clinician and those of society in general. Thus, for a QoL measurement scale, it is unlikely that a single threshold value will be universally accepted as a cut-off point that separates clinically important changes from trivial and unimportant ones. However, many investigators are finding that, for a variety of scales assessing overall QoL and some of its dimensions, changes of between 5% and 10% (that is, between 5 and 10 points on the 1 to 100 scales of the QLQ-C30) are noticed by patients and are regarded by them as “significant changes”.

When QoL is a major outcome measure for a clinical trial, it will be necessary to estimate the required sample size to detect the differences in QoL that are of interest. Methods for doing this are described in the manual of reference data (Fayers et al., 1998b). Before the calculation can be performed, the magnitude of the target difference must be specified. This will be based upon consideration of clinically important differences; prior information regarding plausible treatment differences; and an assessment of the feasibility of accruing the desired number of patients.

Other general references particularly worth consulting are Olschewski et al., 1994; Staquet et al., 1998 and Zee, 1991.

12. ETHICAL ISSUES

Collecting QoL data has ethical implications for both investigator and patient. It is important that patients are fully informed about the reasons for collecting QoL data. They should also be clear about the distinction between their entitlement to a professional concern about their symptoms and QoL, and their participation in “research”.

12.1. Altruism

Participation in QoL studies often has no benefits for the patients themselves but is in the interest of future patients. Their results may be used to improve care and treatment in the future but those who participate in the study often do not benefit personally from participation. The aim of the QoL assessment must therefore be made clear to the patient before inclusion in the study. A separate sheet for informed consent regarding the QoL study may be of great help to the patient.

12.2. Confidentiality & Disclosure

In clinical trials, it is usually recommended that patients' completed questionnaires regarding their QoL are not shown to their physician or other personnel responsible for their treatment. If this is the case, it should be emphasized to the patient at the time of seeking informed consent that it is their responsibility to communicate any problems or symptoms to their doctor. They should be reminded of this throughout the trial. Occasionally patients may indicate such severe levels of symptoms in response to the items in a questionnaire that it should be considered an adverse event. This may give rise to a dilemma between patient safety and patient confidentiality. In this instance the data collector should return to the patient and suggest that they report this symptom to the physician responsible for their treatment. No intervention can be offered to patients who only disclose their symptoms by completing questionnaires (e.g. medication for constipation can only be prescribed if the patient tells their doctor). Patients may also use the opportunity of completing a QoL assessment to talk about and discuss other problems which may be unrelated to their treatment. This may put the person responsible for data collection in a difficult situation. It is important to ensure that the data collector has the opportunity to discuss these issues with others without breaching the patient's confidentiality.

12.3. Eligibility Criteria For Participation

Ideally completing a baseline QoL assessment should be one of the eligibility criteria for a clinical trial. Patients who have good personal reasons for not wishing to participate in a QoL study are then excluded from the study and may be unable to receive the new treatment. A thorough explanation of the aims of the QoL assessments and an assurance of anonymity may overcome these difficulties. Other patients may be unable to read or write but with appropriate assistance may still be able to participate. The patient's decision not to participate, for whatever reason, must be respected. The decision to make participation in the QoL study mandatory must be based on considerations of the nature of the research question e.g. Is it the primary endpoint? Is the study's integrity at risk if QoL assessment is missing?

12.4. Selection Bias

In clinical trials where QoL assessment is relevant, it is important that all eligible patients are included, otherwise the study may not be evaluable and those included may have participated without cause, making the study unethical. However it may also be considered unethical to coerce patients to participate in a QoL study. The problem may be diminished by explaining the rationale for including as many patients as possible in the QoL study, and reinforcing the principle of anonymity.

12.5. End Of Study Assessment

Before the study starts a decision should be made about when QoL assessment will be discontinued. During the trial a number of patients will relapse or a decision will be made that "treatment has failed". The time interval to these events is often one of the endpoints of the trial and no further clinical data is collected, only survival data. The value of QoL assessments beyond these events is debatable. QoL will be overestimated when only those whose treatment is successful remain in the study, but patients should not be burdened with excessive QoL assessments during the last few months of their lives. Some patients may feel discouraged if no one appears to take an interest in their QoL once their treatment has "failed", whilst others may be reluctant to continue participating. The study coordinator and the data manager must be informed about relapses so that proper respect can be shown when sending out reminders to patients with relapse.

12.6. Long Term Follow-up

There are a number of potential problems in studies where long-term postal assessment is planned - especially when the questionnaires are mailed from one central office which relies upon regular updates from the local centers:

- Patients may change address. Not only could this result in missing data but also in a breach of the patient's confidentiality if the new occupant opens the mail.
- Questionnaires may be sent to the home of a patient who has died, which may be distressing for their relatives.
- For some patients who have been in remission for a number of years it becomes distressing to be reminded at regular intervals that they have cancer.
- In all cases it is important that the responsible physician informs the data center about any change in the patient's circumstances on a routine basis and without delay. It is also important that the patients are informed of the long-term nature of the study. The patients can then be considered to have consented to receive questionnaires for a long time.

Even with these measures it is not possible to ensure that questionnaires will never be sent to those who have died during the course of the study. If it does happen then a letter of condolence and an apology should always be sent. Sometimes it may also help to describe the steps that have been taken to avoid the mistake and to explain to the relative the aim of the QoL study and that the patient had consented to participate.

References

Aaronson, N.K. (1995). Quality of Life and clinical trials. *Lance*t, **346**, 1-2.

Aaronson, N.K., Ahmedzai, S., Bergman, B., Bullinger, M., Cull, A., Duez, N.J., Filberti, A., Flechtner, H., Fleishman, S.B., de Haes, J., Kaasa, S., Klee, M., Osoba, D., Razavi, D., Rofe, P., Schraub, S., Sneeuw, K., Sullivan, M., Takeda, F. for EORTC Study Group on Quality of Life. (1993). The European Organisation for Research and Treatment of Cancer QLQ-C30: A Quality-of-Life Instrument for Use in International Clinical Trials in Oncology. *Journal of the National Cancer Institute*, **85**, 365-376.

Bjordal, K., de Graeff, A., Fayers, P.M., Hammerlid, E., van Pottelsberghe, C., Curran, D., Ahlner-Elmqvist, M., Maher, E.J., Meyza, J.W., Brédart, A, Söderholm, A.L., Arrarras, J.J., Feine, J.S., Abendstein, J.S., Morton, R.P., Pignon, T., Huguenin, P., Bottomley, A., Kaasa, S. on behalf of the EORTC Quality of Life Group (2000). A 12 country field study of the EORTC QLQ-C30 (version 3.0) and the head and neck cancer specific module (EORTC QLQ-H&N35) in head and neck patients. *European Journal of Cancer*, **36**, 1796-1807.

Bland, J.M. & Altman, D.G. (1995). Multiple significance tests: The Bonferroni method. *British Medical Journal*, **310**, 170.

Blazeby, J., Sprangers, M., Cull, A., Groenvold, M. & Bottomley, A. on behalf of the Quality of Life Group (2001). Guidelines for *Developing Questionnaire Modules*, (*3rd edition*). Brussels: EORTC.

Cella, D.F., Tulsky, D.S., Gray, G., Sarafian, B., Linn, E., Bonomi, A., Silberman, M., Yellen, S.B., Winicour, P., Brannon, J., Eckberg, K., Lloyd, S., Purl, S., Blendowski, C., Goodman, M., Barnicle, M., Stewart, I., McHale, M., Bonomi, P., Kaplan, E., Taylor, S., Thomas, C.T. & Harris, J. (1993). The Functional Assessment of Cancer Therapy Scale: Development and Validation of the General Measure. *Journal of Clinical Oncology*, **11**, 570-579.

Coates, A., Porzsolt, F. & Osoba, D. (1997). Quality of Life in Oncology Practice: Prognostic Value of EORTC QLQ-C30 Scores in Patients with Advanced Malignancy. *European Journal of Cancer*, **33**, 1025-1030.

Cull, A., Sprangers, M.A.G. , Bjordal, K.& Aaronson, N. on behalf of the EORTC Quality of Life Study Group. (1998). *EORTC Quality of Life Study Group Translation Procedure*. Brussels: EORTC.

Curran, D., Bacchi, M., Hsu Schmitz, S.F., Molenberghs, G. & Sylvester, R.J. (1998a). Identifying The Types of Missingness in Quality of Life Data From Clinical Trials. *Statistics in Medicine*, **17**, 739-756.

Curran, D., Fossa, S., Aaronson, N., Kiebert, G., Keupens, F. & Hall, R. (1997). Baseline quality of life of patients with advanced prostate cancer. *European Journal of Cancer*, **33**, 1809-1814.

Curran, D., Molenberghs, G., Fayers, P. & Machin, D. (1998b). Aspects of incomplete quality of life data in randomised trials: II: Missing forms. *Statistics in Medicine*, **17**, 697-709.

Curran, D., van Dongen, J.P., Aaronson, N., Kiebert, G., Fentiman, I.S., Mignolet, F. & Bartelink, H. (1998c). Quality of life of early breast cancer patients treated with mastectomy or breast conserving procedures: Results of EORTC trial 10801. *European Journal of Cancer*, **34**, 307-314.

de Haes, J.C.J.M., van Knippenberg, F.C.E. & Neijt, J.P. (1990). Measuring psychological and physical distress in cancer patients: Structure and application of the Rotterdam Symptom Checklist. *British Journal of Cancer*, **62**, 1034-1038.

Diggle, P.J., Liang, K.-Y. & Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Oxford University Press: Oxford.

Fayers, P., Curran, D. & Machin, D. (1998a). Incomplete quality of life data in randomized trials: *Missing items. Statistics in Medicine*, **17**, 679-696

Fayers, P., Weeden, S. & Curran, D. on behalf of the EORTC Quality of Life Study Group. (1998b). *EORTC QLQ-C30 Reference Values*. Brussels: EORTC.

Fayers, P.M., Aaronson, N.K., Bjordal, K.,, Groenvold, M., Curran, D. & Bottomley, A. on behalf of EORTC Quality of Life Study Group. (2001). *The EORTC QLQ-C30 Scoring Manual(3rd edition)*. Brussels: EORTC.

Fayers, P.M., Hopwood, P., Harvey, A., Girling, D.J., Machin, D. & Stephens, R. (1997). Quality of life assessment in clinical trials - guidelines and a checklist for protocol writers: The UK Medical Research Council experience. *European Journal of Cancer*, **33**, 20-28.

Fayers, P.M. & Machin, D. (2000). *Quality of life – assessment, analysis and interpretation*. John Wiley & Sons Ltd: Chichester.

Gelber, R.D. & Gelber, S. (1995). *Quality of life assessment in clinical trials*. Recent advances in clinical trial design and analysis. Kluwer Academic Publishers: Boston.

Goldstein, H. (1995). *Multilevel Statistical Models*. Edward Arnold: London.

Gotay, C.C., Korn, E.L., McCabe, M.S., Moore, T.D. & Cheson, B.D. (1992). Quality-of-life assessment in cancer treatment protocols: Research issues in protocol development. *Journal of the National Cancer Institute*, **84**, 575-579.

Hand, D.J. & Crowder, M. (1996). *Practical Longitudinal Data Analysis*. Chapman & Hall: London.

Hjermstad, M.J., Fayers, P.M., Bjordal, K. & Kaasa, S. (1998). Health-related quality of life in the general Norwegian population assessed by the EORTC core Quality of Life Questionnaire - The QLQ-C30 (+3). *Journal of Clinical Oncology*, **16**, 1188-1196.

King, M.T. (1996). The interpretation of scores from the EORTC quality of life questionnaire QLQ-C30. *Quality of Life Research*, **5**, 555-567.

Lindsey, J.K. (1993). *Models for Repeated Measurements*. Oxford University Press: Oxford.

Matthews, J.N.S., Altman, D.G., Campbell, M.J. & Royston, P. (1990). Analysis of serial measurements in medical research. *British Medical Journal*, **300**, 230-235.

Olschewski, M., Schulgen, G., Schumacher, M. & Altman, D.G. (1994). Quality of life assessment in clinical cancer research. *British Journal of Cancer*, **70**, 1-5.

Osoba, D. (1992). The Quality of Life Committee of the Clinical Trials Group of the National Cancer Institute of Canada: Organisation and Functions. *Quality of Life Research*, **1**, 211-218.

Osoba, D., Pater, J.L. & Zee, B. (1994). Effective anti-emetic therapy improves quality of life (QoL) after moderately emetogenic chemotherapy (MEC). *Quality of Life Research*, **4**, 467-468.

Osoba, D., Rodrigues, G., Myles, J., Zee, B. & Pater, J. (1998). Interpreting the significance of changes in health-related quality of life scores. *Journal of Clinical Oncology*, **16**, 139-144.

Rosendahl, K.I., Curran, D., Kiebert, G., Cole, B., Weeks, J.C., Denis, L.J. & Hall, R.R. (1997). A quality-adjusted survival (Q-Twist) analysis of EORTC trial 30853 comparing maximal androgen blockade (MAB) with orchidectomy in patients with metastatic prostate cancer. In *American Society of Clinical Oncology* (ASCO), Vol. 16. pp. 312 (A1113).

Schipper, H., Clinch, J., McMurray, A. & Levitt, M. (1984). Measuring the quality of life of cancer patients: The Functional Living Index-Cancer: Development and Validation. *Journal of Clinical Oncology*, **2**, 472-483.

Slevin, M.L., Stubbs, L., Plant, H.J., Wilson, P., Gregory, W.M., Armes, P.J. & Downer, S.M. (1990). Attitudes to chemotherapy: comparing views of patients with cancer with those of doctors, nurses and general public. *British Medical Journal*, **300**, 1458 -1460.

Sneeuw, K.C.A., Aaronson, N.K., Sprangers, M.A.G., Detmar, S.B., Wever, L.D.V., Schornagel, J.H. (1997). The value of caregiver ratings in evaluating the quality of life of patients with cancer. *Journal of Clinical Oncology*, **15**,1206-1217.

Sneeuw, K.C.A., Aaronson, N.K., Sprangers, M.A.G., Detmar, S.B., Wever, L.D.V., Schornagel, J.H. (1998). Comparison of patient and proxy EORTC QLQ-C30 ratings in assessing the quality of life of cancer patients. *Journal of Clinical Epidemiology*, **51**, 617-631.

Staquet, M.J., Hays, R.D. & Fayers, P.M. (1998). *Quality of Life Assessment in Clinical Trials: Methods and Practice.* Oxford University Press: Oxford.

Stephens, R.J., Hopwood, P., Girling, D.J. & Machin, D. (1997). Randomized trials with quality of life endpoints: Are doctors' ratings of patients' physical symptoms interchangeable with patients' self-ratings? *Quality of Life Research*, **6**, 225-236.

Tandon, P.K. (1990). Applications of global statistics in analysing quality of life data. *Statistics of Medicine*, **9**, 819-827.

Tannock, I.F., Osoba, D. & Stockler, M.R. (1996). Chemotherapy with mitoxantrone plus prednisone or prednisone alone for symptomatic hormone-resistant prostate cancer: A Canadian randomized trial with palliative endpoints. *Journal of Clinical Oncology*, **14**, 1756-1764.

Troxel, A.B., Fairclough, D.L., Curran, D. & Hahn, E.A. (1998). Statistical Analysis of Quality of Life Data in Cancer Clinical Trials. *Statistics in Medicine*, **17**, 653-666.

Vachalec, S., Bjordal, K., Bottomley, A., Blazeby, J., Flechtner, H. & Ruyskart, P. on behalf of the EORTC Quality of Life Group (2001). *Item Bank Guidelines*. Brussels: EORTC.

Ware, J.E. (1987). Standards for validating health measures: definition and content. *Journal of Chronic Diseases*, **40**, 473.

Zee, B. (1991). In: In *Effects of Cancer on Quality of Life*, Osoba, D. (ed). CRC Press.

Appendices

1. QLQ-C30 v3

2. List of available QLQ-C30 translations

3. QLU staff and contact addresses

4. List of available modules, translations and contact names

5. List of joint scientific committee members and contact addresses

6. Codes for missing data

7. List of all EORTC studies assessing QoL

8. EORTC Guidelines for administration of QoL questionnaires

Quality of Life Unit,
EORTC Data Center
Avenue E. Mounier 83 – Bte 11,
1200 Brussels
BELGIUM

Tel: +32 2 774 1680
Fax: +32 2 779 4568

D/2002/6136/003
ISBN 2-930064-27-7