Volume 8 • Number 2 • 2005
VALUE IN HEALTH

Using Multiple Anchor- and Distribution-Based Estimates to Evaluate Clinically Meaningful Change on the Functional Assessment of Cancer Therapy-Biologic Response Modifiers (FACT-BRM) Instrument

Kathleen J. Yost, PhD,[1] Mark V. Sorensen, PhD,[2] Elizabeth A. Hahn, MA,[1] G. Alastair Glendenning, MSc,[3] Ari Gnanasakthy, MBA, MSc,[4] David Cella, PhD[1]

[1]Center on Outcomes, Research and Education (CORE), Evanston, IL, USA; [2]Department of Anthropology, University of North Carolina, Chapel Hill, NC, USA; [3]Novartis AG, Horsham, Sussex, UK; [4]Novartis Pharmaceutical Corporation, East Hanover, NJ, USA

ABSTRACT

Objective: The interpretation of health-related quality of life (HRQL) data from clinical trials can be enhanced by understanding the degree of change in HRQL scores that is considered meaningful. Our objectives were to combine distribution-based and two anchor-based approaches to identify minimally important differences (MIDs) for the 27-item Trial Outcome Index (TOI), the seven-item Social Well-Being (SWB) subscale, and the six-item Emotional Well-being (EWB) subscale from the Functional Assessment of Cancer Therapy-Biological Response Modifiers (FACT-BRM) instrument.

Methods: Distribution-based MIDs were based on the standard error of measurement. Anchor-based approaches utilized patient-reported global rating of change (GRC) and change in physician-reported performance status rating (PSR). Correlations and weighted kappa statistics were used to assess association and agreement between the two anchors. FACT-BRM changes were evaluated for three time periods: baseline to month 1, month 2 to month 3, and month 5 to month 6.

Results: Association between GRC and change in PSR was poor. Correlation between the anchors and HRQL change scores was largest at month 1 and decreased through month 6. Combining results from all approaches, the MIDs identified were 5–8 points for the TOI, 2 points for the SWB subscale, and 2–3 points for the EWB subscale.

Conclusions: We combined patient-reported estimates, physician-reported estimates, and distribution-based estimates to derive MIDs for HRQL outcomes from the FACT-BRM. These results will enable interpretation of treatment group effects in a clinical trial setting, and they can be used to estimate sample size or power when designing future studies.

Keywords: clinical significance, clinical trials, health-related quality of life, minimally important difference.

Introduction

Health-related quality of life (HRQL) is a multidimensional construct that refers to the extent to which one's usual or expected physical, emotional, and social well-being are affected by a medical condition or its treatment [1]. By its very nature, HRQL is subjective. Interest in evaluating HRQL outcomes in clinical trials continues to increase with greater emphasis on measuring patient-reported outcomes, evaluating what is an important change from the patient's perspective [2], and including patient-reported outcomes as primary end points in clinical trials [3]. Providing HRQL information to patients becomes increasingly important for making treatment decisions that may affect length of survival, functional status, or pain and symptom management [4,5].

A number of issues need to be considered when including HRQL measures in clinical trials. Use of rigorously validated HRQL instruments is critical, because it is a need to ensure that results are mean-

Address correspondence to: Kathleen J. Yost, Center on Outcomes, Research and Education (CORE), Evanston Northwestern Healthcare, 1001 University Place, Suite 100, Evanston, IL 60201, USA. E-mail: kyost@enh.org

Portions of this article were presented at the Assessing Clinical Significance for Quality of Life Measures in Oncology Research: State-of-the-Science Quality of Life Symposium, Rochester, MN, April 2002, and at the Drug Information Association 9th Annual Symposium for Quality of Life Evaluation, Hilton Head Island, SC, USA, March 2002.

ingful to both clinicians and patients. Logistical and methodological issues to consider include the timing of assessments, selection of relevant HRQL domains, sample size estimation, and specification of hypotheses to be tested. Defining the meaningfulness of change in HRQL scores is critical for determining appropriate sample sizes, interpreting treatment group results, and understanding change over time. Determination of a clinically meaningful change is challenging because HRQL measures lack a "gold standard" against which to quantify meaningful change.

Meaningful change in HRQL can be assessed using distribution- or anchor-based methods [6,7]. Distribution-based measures rely on the statistical distributions of HRQL data, and include effect size measures [8–10], the standard error of measurement (SEM) [11,12], the responsiveness index [13] and the reliable change index [14]. Anchor-based approaches involve comparing changes in HRQL scores to patient-reported assessments of change over time [15,16] or to clinically relevant measures such as performance status or response to treatment [17,18]. Jaeschke and colleagues [19] developed an anchor-based approach using a global rating of change (GRC) to estimate what they describe as the minimal clinically important difference (MCID) in HRQL change scores, defined as "the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management [19]." The terminology for meaningful change is varied [20] and lacks consensus [7,21]. We employ the more commonly used term of minimally important difference (MID) herein.

The GRC is a retrospective assessment of perceived change; that is, patients are asked to think back to a previous time point and state whether they have experienced any change in a domain of health and/or HRQL from that time point to the present. Response options typically range from "very much worse" through "about the same" to "very much better." We and others have applied the Jaeschke et al. method to the analysis of HRQL [15,19], including the analyses of change in HRQL scores in a clinical trial [16]. Serial, or prospective anchors, have also been used to identify MIDs [18,22,23]. For a prospective anchor, a patient is rated with respect to the clinical anchor at two time points; for example, once at baseline before commencing treatment and again 1 month later. The change in the anchor over these two time points can be used to assess the clinical meaningfulness of change in HRQL scores over the same time.

The Functional Assessment of Cancer Therapy—General (FACT-G) [24] is a 27-item instrument developed to measure HRQL in cancer patients. The FACT-G measures the general domains of HRQL, including physical well-being (PWB, seven items), functional well-being (FWB, seven items), social/family well-being (SWB, seven items) and emotional well-being (EWB, six items). A common practice is to supplement a general HRQL instrument with additional items that capture information specific to a particular disease or treatment [25]. This approach was taken when developing the FACT-Biological Response Modifiers (FACT-BRM) instrument, which supplements the FACT-G with 13 disease- and treatment-specific items addressing concerns commonly experienced by cancer patients undergoing treatment with biological response modifiers [26–28]. These 13 items form two subscales, physical BRM and emotional/cognitive BRM, or they can be combined into a single subscale. The Trial Outcome Index (TOI) is a 27-item summary measure of physical function and well-being, and it is derived by summing the PWB, FWB, and BRM subscale scores. TOI scales have been derived for other instruments in the Functional Assessment of Chronic Illness Therapy (FACIT) measurement system and have proven to be valuable summary measures in clinical trials [18,22,29]. The Cronbach's alpha for the 27-item TOI from the FACT-BRM is very high (0.91–0.93) [27], which supports combining these items to create a single scale. More information about the FACT-BRM and other FACIT instruments is available at http://www.facit.org

The purpose of the present study was to identify MIDs for the 27-item TOI, the seven-item SWB subscale, and the six-item EWB subscale. We illustrate an approach for identifying MIDs that combines distribution-based estimates and two types of anchor-based estimates.

Methods

Participants were newly diagnosed patients with chronic phase chronic myelogenous leukemia participating in the International Randomized Interferon versus ST1571 (IRIS) Study, a prospective multicenter open-label Phase III randomized study [27,30]. A total of 1106 patients were randomized to either oral imatinib (STI571) 400 mg daily or subcutaneous interferon-alpha (IFNα) plus low-dose Ara-C (IFNα + LDAC). A substudy was conducted during the first 6 months of treatment in a subset of 200 patients in the United States. The principal investigator of the substudy (A.G.) designed a GRC questionnaire, described in detail below, which was completed by 175 (87.5%) of these patients.

HRQL Assessment

HRQL was measured monthly during the first 6 months of the trial using the FACT-BRM [27,28]. The GRC substudy was designed to identify an MID for the TOI, the primary HRQL end point in the IRIS trial. We also wanted to identify MIDs for the SWB and EWB subscales, which were considered secondary HRQL end points. Scoring of the FACT-BRM was performed as described in the FACIT manual [26].

Patient-Reported GRC

A self-administered GRC questionnaire was developed based on the methods established by Jaeschke et al. [19]. The meaning of each subscale (PWB, SWB, EWB, FWB, BRM) as well as overall HRQL was briefly described and then patients rated the degree of change they had experienced in each of these domains since the last time they completed the FACT-BRM. Thus, there was a GRC rating for each of the subscales (GRC_{PWB}, GRC_{SWB}, GRC_{EWB}, GRC_{FWB}, GRC_{BRM}) and overall HRQL (GRC_{Total}). Patients reported change on a five-point scale ranging from 1 = "a lot better" to 5 = "a lot worse." Patients provided these ratings at months 1, 3, and 6. The GRC measured at month 1 reflected patients' perceived change from baseline to month 1, whereas GRC measured at month 3 reflected perceived change from month 2, and GRC at month 6 reflected perceived change from month 5. To evaluate a GRC rating corresponding to the TOI (GRC_{TOI}), the mean of the GRC_{PWB}, GRC_{FWB}, and GRC_{BRM} was computed. Mean scores were then rounded to the nearest integer to correspond to the ratings on the GRC questionnaire. Because of small sample sizes in some of the GRC categories, patients were further classified into three categories by combining the "a lot better" and "somewhat better" categories into one category ("better"), and by combining the "somewhat worse" and "a lot worse" categories into a single category ("worse").

Physician-Reported Performance Status Rating (PSR)

The Eastern Cooperative Oncology Group (ECOG) PSR is an assessment of current functional status often used in clinical trials. PSR grades range from 0 to 4, where 0 = fully ambulatory without symptoms, 1 = fully ambulatory with symptoms, 2 = requiring bed rest less than half of the waking day, 3 = requiring bed rest more than half of the waking day, and 4 = bedridden. Only patients with PSR grades of 0, 1, or 2 were eligible for the IRIS clinical trial. Physician-determined PSR was recorded at each monthly visit; thus, PSR can be considered a prospective anchor. We computed change in PSR for three time intervals: month 1 minus baseline, month 3 minus month 2, and month 6 minus month 5. Patients whose PSR declined by 1 or more points were classified as "worse," patients whose PSR did not change were classified as "same," and patients who improved by 1 or more points were classified as "better."

Statistical Analysis

Descriptive statistics. Internal consistency reliability of the FACT-BRM was assessed using Cronbach's coefficient alpha. To evaluate the concordance between patient-reported and physician-reported estimates of change, Spearman rank correlation coefficients were computed for PSR change scores and GRC ratings, both defined in three categories of "worse," "same," and "better." Because PSR change and GRC_{TOI} are both predominantly measures of change in physical functioning, we expected the correlation between PSR change and GRC_{TOI} to be larger than correlations between PSR change and either GRC_{SWB} or GRC_{EWB}. Weighted kappa statistics were also used to assess agreement between PSR change and GRC defined in three categories.

Distribution-based approach. Meaningful change was defined as one SEM and was computed at baseline using the following formula: $SEM = \sigma\sqrt{1 - r_{xx}}$ where σ is the standard deviation of the subscale and r_{xx} is the reliability of the subscale, measured here as Cronbach's alpha [11]. Although other multiples of the SEM, including 1.96 SEM [14] and 2.77 SEM [31], have been used to identify meaningful change, we chose 1.0 SEM because this value has been specifically linked to the MID in HRQL [11,12]. Effect sizes were computed by dividing the SEM by the standard deviation of the baseline score.

Anchor-based approach. Two anchors were used to assess meaningful change in HRQL: 1) patient-reported GRC; and 2) physician-reported PSR. Changes in HRQL scores were computed within each GRC category (worse, same, better). Effect sizes for change in HRQL scores at month 1 were computed by dividing the mean change score from baseline to month 1 by the standard deviation of the baseline score. The HRQL mean change score from month 2 to month 3 was divided by the month 2 standard deviation and the mean change score from month 5 to month 6 was divided by the month 5 standard deviation to derive effect sizes for those assessments. Patients were also grouped according to PSR change category, and mean change in HRQL scores and effect sizes were computed within each category as described above. An effect size of 0.2 was considered small, 0.5 was moderate and 0.8 was large [8]. FACT-BRM change scores should increase monotonically in line with moving from the categories "worse" to "better" on the GRC and PSR change. For example, changes in TOI scores for the "worse" GRC_{TOI} category should be negative, changes in the "about the same" GRC_{TOI} category should be near zero, and changes for the "better" GRC_{TOI} category should be positive.

Crosby et al. [7] suggest that the usefulness of an anchor depends on the correlation between the HRQL change score and the anchor. We used statistical significance ($P < 0.05$) of a Spearman correlation coefficient to assess the usefulness of anchors. In this study, if the correlation between the anchor and the change in HRQL score was not positive and not statistically significant, the anchor was considered inappropriate for assessing change in HRQL, and the HRQL change score was not included in the determination of the MID. We anticipated the correlations between the GRC and HRQL change to be moderate to large. Because the PSR change is a measure of change in physical functioning, we anticipated the correlation with change in TOI scores to be moderate to large and the correlations with change in SWB and EWB scores to be small.

Clinically meaningful change. No single method for identifying MIDs is sufficient; therefore, it has been recommended that multiple strategies be used simultaneously [32,33]. Furthermore, because MIDs may vary slightly across patient groups, reporting a range of plausible MIDs rather than a single number has been recommended [32,33]. Mean change scores with large effect sizes (> 0.8) [8] were not included in the determination of the MID because, while clinically meaningful, these changes are too large to be considered MIDs. We summarized the anchor-based MID estimates using medians and interquartile ranges to facilitate identifying a range of plausible MIDs. The interquartile range has been recommended for providing reasonable bounds around the MID because it is robust to possibly asymmetric distributions of MID estimates [34].

Meaningful declines in FACT-G scores based on the GRC anchor were larger on average than meaningful improvements [15]. This trend was also observed for the FACT-Colorectal based on two patient-reported anchors: general health and bowel function [23]. Meaningful declines in FACT scores based on change in physician- and patient-reported PSR, however, have tended to be smaller on average than meaningful improvements [18,35]. Thus, the asymmetric nature of meaningful change in HRQL scores may depend on the anchor used. Others have observed a trend in HRQL change scores suggesting declines are larger than improvements [36,37], but a recent review of MID research concluded there was insufficient evidence to support the assertion that MIDs for worsening states were larger than those for improving states [20]. We computed MIDs for all patients combined, but we also present plausible MIDs separately for patients who decline and improve as secondary or exploratory findings.

Based on results from an expert panel [38], the smallest clinically important difference of a scale should be larger than its intrinsic variability (defined here as one SEM). Therefore, we ensured that the ranges of MIDs identified included the next possible HRQL score larger than the SEM. To enhance interpretation of MIDs across scales with different numbers of items, we described the MIDs in terms of average change per item. For example, an MID of 2 points for a seven-item scale corresponds to a change of 2/7 = 0.29 points per item.

Results

Baseline demographic and clinical characteristics of the substudy population are presented in Table 1, and baseline descriptive statistics for the FACT-BRM scores are presented in Table 2. The characteristics of patients in this substudy were similar to those for all patients enrolled in the IRIS trial [39]. The mean SWB score for this substudy was higher than the general US population norm, but the mean EWB score was lower than the norm [40]. There are currently no normative data for the TOI from the FACT-BRM instrument.

Correlations between change in PSR and GRC_{TOI}, GRC_{SWB}, and GRC_{EWB} were highest at month 1 (Table 3). These correlations decreased and remained fairly consistent at months 3 and 6. As expected, the correlations between PSR change and GRC_{TOI} were larger than the correlations between PSR change and GRC_{SWB} or GRC_{EWB} at all

Table 1 Baseline demographic and clinical characteristics

	US patients N = 175
Age (year)	
Median (range)	52 (18–70)
≥60, no. (%)	28 (18)
Sex, no. (%)	
Male	107 (61.1)
ECOG performance status rating, no. (%)	
0 = Fully ambulatory without symptoms	136 (78.2)
1 = Fully ambulatory with symptoms	35 (20.1)
2 = less than 50% bed rest during waking day	3 (1.7)
Interval from diagnosis (month)	
Median (range)	1.7 (0–7.3)
Sokal risk group, no. (%)	
Low	49 (28.0)
Intermediate	44 (25.1)
High	13 (7.4)
Unknown	69 (39.4)
Hasford risk group, no (%)	
Low	61 (34.9)
Intermediate	30 (17.1)
High	21 (12.0)
Unknown	63 (36.0)

time points. Overall agreement between PSR change and GRC as measured by a weighted kappa statistic was poor (Table 3). Agreement was better at the month 1 assessment than at months 3 and 6, and it was better for PSR change versus GRC$_{TOI}$ than for PSR change versus GRC$_{SWB}$ or GRC$_{EWB}$.

Distribution-Based Analysis

The estimates of one SEM were 4.2 points for the TOI, 1.9 points for the SWB, and 2.2 points for the EWB (Table 2). Effect sizes associated with these SEMs were 0.25 for the TOI, 0.48 for the SWB, and 0.50 for the EWB, which were considered small to moderate [8].

Anchor-Based Analyses

Mean changes in TOI scores are presented in Table 4. TOI change scores were asymmetric in that the absolute values of the mean change scores for the "worse" category were markedly larger than changes in the "better" category for all but PSR change at months 3 and 6. TOI change scores in the "about the same" group were near zero for the month 3 and month 6 assessments, as expected.

Table 3 Spearman correlations and weighted kappa statistics for PSR change versus GRC$_{TOI}$, GRC$_{SWB}$, GRC$_{EWB}$

	GRC$_{TOI}$	GRC$_{SWB}$	GRC$_{EWB}$
Correlation			
Month 1	0.37	0.24	0.23
Month 3	0.25	0.13	0.09
Month 6	0.25	0.11	0.09
Kappa statistic			
Month 1	0.28	0.13	0.15
Month 3	0.12	0.00	0.00
Month 6	0.18	0.12	0.05

PSR change, change in physician-reported performance status rating; GRC$_{EWB}$, patient-reported global rating of change in emotional well-being; GRC$_{SWB}$, patient-reported global rating of change in social well-being; GRC$_{TOI}$, patient-reported global rating of change in physical functioning and well-being.

Table 4 Change in TOI scores by patient-reported GRC ratings and by change in physician-reported PSR

	n	TOI change score mean (SD)	Overall SD*	Effect size[†]
GRC$_{TOI}$ groups				
Month 1				
Worse	57	−28.2 (17.1)[‡]		−1.76
About the same	61	−5.6 (12.3)	16.0	−0.35
Better	43	5.7 (14.6)		0.31
Month 3				
Worse	23	−8.9 (11.1)		−0.42
About the same	71	0.1 (10.7)	21.1	0.00
Better	52	6.9 (13.5)		0.33
Month 6				
Worse	18	−8.8 (9.0)		−0.43
About the same	80	0.9 (7.9)	20.3	0.04
Better	53	3.6 (12.2)		0.18
PSR change groups				
Month 1				
Worse	39	−27.5 (21.3)[‡]		−1.71
About the same	103	−6.3 (16.1)	16.1	−0.39
Better	15	4.3 (15.9)		0.27
Month 3				
Worse	22	−5.5 (10.9)		−0.26
About the same	108	0.7 (10.7)	21.0	0.03
Better	16	9.9 (21.4)		0.47
Month 6				
Worse	11	−7.2 (8.6)		−0.36
About the same	117	0.9 (8.9)	20.2	0.05
Better	16	6.2 (15.2)		0.31

*SD at baseline, month 2 or month 5.

[†] Effect size calculated as the change score divided by the overall SD.

[‡] This estimate of change was not considered in the determination of the MID because the effect size is greater than 0.8.

GRC$_{TOI}$, patient-reported global rating of change in physical functioning and well-being; PSR, physician-reported performance status rating. See text for description of change; TOI, trial outcomes index.

Table 2 Baseline HRQL scores

FACT scale/subscale	Range	Mean (SD)	Cronbach's alpha	SEM	US general population norm mean (SD) [40]
Trial Outcome Index (TOI, 27-items)	22–108	84.9 (16.5)	0.94	4.2	NA
Social Well-Being (SWB, 7-items)	4–28	24.3 (4.0)	0.77	1.9	19.1 (6.8)
Emotional Well-Being (EWB, 6-items)	4–24	17.8 (4.4)	0.76	2.2	19.9 (4.8)

SEM, one standard error of measurement; NA, not available.

Nevertheless, patients classified as "about the same" according to their GRC_{TOI} or PSR change at month 1 experienced declines in their TOI scores of –5.6 for GRC_{TOI} and –6.3 for PSR change, which correspond to small to moderate effect sizes of –0.35 and –0.39, respectively.

The results of the anchor-based analysis for the SWB subscale are in Table 5. Absolute change scores for patients classified as "worse" based on GRC_{SWB} tended to be larger than change scores for patients classified as "better"; whereas changes for declines versus improvements based on PSR change were essentially equivalent. SWB change scores were similar for the "about the same" and "better" categories of GRC_{SWB} at both the month 3 and 6 assessments. This absence of a monotonic increase in SWB change scores across categories of GRC_{SWB} at months 3 and 6 indicates that self-reported GRC_{SWB} did not reflect change in the SWB subscale at those assessments. SWB was slightly more responsive to PSR change as indicated by negative change scores in the "worse" category and positive change scores in the "better" category. The GRC_{SWB} anchor suffered from small sample sizes in the "worse" category at all assessments.

Table 5 Change in SWB scores by patient-reported GRC ratings and by change in physician-reported PSR

	n	SWB change score mean (SD)	Overall SD*	Effect size[†]
GRC_{SWB} groups				
Month 1				
Worse	13	−3.8 (4.7)[‡]		−0.96
About the Same	99	−1.4 (2.9)	3.9	−0.35
Better	47	0.4 (4.9)		0.10
Month 3				
Worse	8	−2.8 (4.3)[§]		−0.64
About the Same	97	−0.03 (3.0)	4.4	0.00
Better	45	−0.06 (2.7)[§]		0.02
Month 6				
Worse	8	−1.6 (3.2)[§]		−0.38
About the Same	92	−0.1 (3.1)	4.3	−0.02
Better	45	−0.03 (2.3)[§]		0.00
PSR change groups				
Month 1				
Worse	40	−2.2 (3.6)		−0.56
About the Same	104	−1.0 (3.1)	3.9	−0.26
Better	15	2.4 (7.1)		−0.62
Month 3				
Worse	22	−0.3 (1.7)[§]		−0.07
About the Same	112	−0.4 (3.1)	4.6	−0.09
Better	16	0.2 (2.3)[§]		0.04
Month 6				
Worse	11	−1.6 (3.2)		−0.33
About the Same	118	−0.2 (2.8)	4.9	−0.04
Better	16	1.4 (3.0)		0.29

*SD at baseline, month 2 or month 5.

[†] Effect size calculated as the change score divided by the overall SD.

[‡] This estimate of change was not considered in the determination of the MID because the effect size is greater than 0.8.

[§] This estimate of change was not considered in the determination of the MID because the SWB change scores were not significantly correlated with the anchor.

GRC_{SWB}, patient-reported global rating of change in social well-being; PSR: physician-reported performance status rating. See text for description of change; SWB, Social Well-Being subscale.

Table 6 Change in EWB scores by patient-reported GRC ratings and by change in physician-reported PSR

	n	EWB change score mean (SD)	Overall SD*	Effect size[†]
GRC_{EWB} groups				
Month 1				
Worse	28	−2.1 (5.1)		−0.49
About the same	76	0.8 (3.5)	4.3	0.19
Better	60	1.9 (3.8)		0.45
Month 3				
Worse	19	−1.1 (3.1)		−0.27
About the same	82	−0.4 (3.0)	4.1	−0.10
Better	48	0.7 (2.8)		0.12
Month 6				
Worse	18	−0.2 (3.9)[‡]		−0.05
About the same	75	−0.1 (2.3)	3.9	−0.03
Better	58	0.7 (3.2)[‡]		0.18
PSR change groups				
Month 1				
Worse	39	−0.6 (4.7)		−0.07
About the same	105	0.9 (3.8)	4.4	−0.10
Better	15	2.5 (3.7)		0.26
Month 3				
Worse	22	−0.3 (2.3)[‡]		−0.07
About the same	111	−0.4 (3.0)	4.2	−0.10
Better	16	1.1 (2.8)[‡]		0.26
Month 6				
Worse	11	−1.7 (3.6)		−0.38
About the same	118	0.3 (2.4)	4.5	0.07
Better	16	1.6 (4.1)		0.35

*SD at baseline, month 2 or month 5.

[†] Effect size calculated as the change score divided by the overall SD.

[‡] This estimate of change was not considered in the determination of the MID because the EWB change scores were not significantly correlated with the anchor.

EWB, emotional well-being subscale; GRC_{EWB}, patient-reported global rating of change in emotional well-being; PSR: physician-reported performance status rating. See text for description of change.

Table 6 contains the results of the anchor-based analysis for EWB. For this subscale, change scores for declines were essentially the same magnitude as those for improvements based on the GRC_{EWB} at months 1 and 3 and based on PSR change at months 6. Improvements were larger than declines for the GRC_{EWB} at month 6 and for PSR change at months 1 and 3. There was evidence of a lack of monotonic increase in EWB change score across the anchors for the GRC_{EWB} at month 6 and the PSR change at month 3.

Clinically Meaningful Change

Correlations between FACT-BRM scores and the anchors are summarized in Table 7. The correlations between change in TOI and both anchors (GRC_{TOI} and change in PSR) were statistically significant at all assessments; therefore, they were considered appropriate anchors for measuring

Table 7 Spearman rank correlations between FACT-BRM scores and GRC or PSR change

FACT-BRM scale/ subscale	Anchor (assessment period)	Correlations		
		Pre*	Post†	Change‡
TOI	GRC_{TOI} (month 1)	0.00§	0.59	0.68
	GRC_{TOI} (month 3)	0.28	0.54	0.39
	GRC_{TOI} (month 6)	0.21	0.33	0.24
SWB	GRC_{SWB} (month 1)	0.01§	0.23	0.23
	GRC_{SWB} (month 3)	0.17	0.21	0.11§
	GRC_{SWB} (month 6)	0.32	0.29	0.09§
EWB	GRC_{EWB} (month 1)	0.18	0.45	0.33
	GRC_{EWB} (month 3)	0.18	0.29	0.22
	GRC_{EWB} (month 6)	0.25	0.36	0.08§
TOI	PSR Change (month 1)	–0.06§	0.35	0.48
	PSR Change (month 3)	0.02§	0.18	0.25
	PSR Change (month 6)	0.08§	0.19	0.25
SWB	PSR Change (month 1)	–0.11§	0.09§	0.26
	PSR Change (month 3)	–0.04§	–0.02§	0.05§
	PSR Change (month 6)	0.03§	0.13§	0.22
EWB	PSR Change (month 1)	–0.03§	0.15§	0.19
	PSR Change (month 3)	–0.04§	0.02§	0.13§
	PSR Change (month 6)	–0.02§	0.11§	0.21

*Pre, correlation between anchor and HRQL score measured at Baseline, month 2, or month 5.

†Post, correlation between anchor and HRQL score measured at month 1, month 3, or month 6.

‡Change, correlation between anchor and HRQL change score.

§Not statistically significant. All other $P < 0.05$.

EWB, emotional well-being subscale; FACT-BRM, Functional Assessment of Cancer Therapy-Biological Response Modifiers; GRC, patient-reported global rating of change; PSR change, change in physician-reported performance status rating. See text for description of change; SWB, social/family well-being subscale; TOI, trial outcomes index.

change in TOI. Mean changes in the "worse" category at the month 1 assessment were –28.2 points for the GRC (effect size –1.76) and –27.5 points for the PSR change (effect size –1.71; see Table 4). While these changes are clinically meaningful, they are too large to be considered a *minimal* criterion for assessing meaningful change; thus, these values were not included in the determination of an MID for the TOI scale. The remaining absolute anchor-based change scores ranged from 3.6 to 9.9 points (Table 4), with a median of 6.6 and an interquartile range of 5.6–8.4. The distribution-based estimate (i.e., SEM) was 4.2 points (Table 2). Based on these estimates, we recommend that changes in TOI scores between 5 and 8 points be considered minimally clinically important. The lower end of this range is larger than the SEM, which meets the recommendation that the MID should be larger than the inherent variability of the scale [38]. Evaluating estimates separately for decliners and improvers suggests MIDs of 8 points for decliners and 6 points for improvers.

The correlations between change in SWB scores and GRC_{SWB} were not statistically significant at months 3 and 6, and for PSR change, correlations were not statistically significant at month 3 (Table 7), indicating that these anchors were not suitable for assessing clinically meaningful change in the SWB subscale at those times. The mean month 1 change score in the "worse" category of GRC_{SWB} was –3.8 and had an effect size of –0.96, which exceeds our effect size criterion of 0.8 for identifying *minimally* important changes. Therefore, this estimate was not included in the determination of the MID. Absolute values of the remaining estimates that were considered ranged from 0.4 to 2.4 with a median of 1.6 and an interquartile range of 1.4–2.2. The distribution-based estimate (SEM) is 1.9. Combining the anchor- and distribution-based results suggests that 2 points is a plausible MID for SWB subscale scores. There was inconsistent evidence that SWB change scores in the "worse" category were larger than in the "better" category; therefore, MIDs were not evaluated separately for declines versus improvements.

The correlation between EWB change scores and GRC_{EWB} was not statistically significant at month 6, and it was not significant between EWB change scores and PSR change at month 3 (Table 7); thus, results from these assessments were not considered appropriate for assessing change in EWB. The anchor-based MID estimates ranged in absolute value from 0.6 to 2.5 points with a median of 1.7 and an interquartile range of 1.0–2.0 (see Table 6). The distribution-based estimate was 2.2 points (Table 2). The distribution- and anchor-based estimates cluster around 2 points. Nevertheless, because the MID range should include the next score larger than the SEM, our recommended range of the MID for the EWB subscale is 2–3 points. The data in Table 6 do not support establishing separate MIDs for declines versus improvements.

Discussion

We combined distribution-based and two anchor-based methods to identify the magnitude of a change in three FACT-BRM end points that could be considered minimally clinically meaningful. Recommended MIDs were 5–8 points for the TOI (0.19–0.30 points per item), 2 points for the SWB (0.29 points per item), and 2–3 points for the EWB (0.33–0.50 points per item). The magnitudes of the MIDs differ because the scales and subscales are comprised of different numbers of items with different score ranges and different distributional and psychometric properties. Although we also explored the possibility of separate MIDs for the

TOI for declines (8 points) and for improvements (6 points), TOI declines were not consistently and markedly larger than improvements at all assessments evaluated in this study. Thus, at this time we recommend that the MID range based on all patients combined (i.e., 5–8 points) be used to interpret TOI scores.

Effect sizes for 91% of the TOI estimates, 78% of the SWB estimates, and 50% of the EWB estimates considered in the determination of MIDs were between 0.20 and 0.50, which are interpreted as small to medium effects [8]. Therefore, we are confident that the recommended MIDs do, in fact, approximate the *minimum* change score one would consider meaningful in the selected FACT-BRM end points.

There is growing evidence that MIDs for FACIT scales and subscales are fairly stable across patient populations. The recommended MID for the 27-item TOI scale ranged from 5–8 points, which is 0.19–0.30 points per item. This is very consistent with findings from previous research on MIDs for TOI scales from other FACIT instruments such as the FACT-Lung [22], FACT-Breast [18], and FACT-Colorectal [23], for which MIDs for TOIs consistently ranged from 0.19–0.29 points per item [23]. MIDs of 2–3 points for the seven-item PWB and FWB subscales have been previously reported [15], which correspond to a range of 0.29–0.43 points per item. MIDs for cancer-specific subscales of FACIT instruments range from 0.22–0.43 points per item [23]. Our findings in the present study for the SWB (0.29 points per item) and EWB (0.33–0.50 points per item) are consistent with these previous reports. Based on these findings, we are confident that the MIDs for TOI, SWB, and EWB identified in this clinical trial are appropriate for interpreting HRQL data from other patient samples.

We identified ranges of MIDs for the TOI and EWB, as recommended [32,33]. The data in this study did not support a range of estimates for the SWB subscale; the anchor-based estimates clustered around 2 points and the distribution-based estimate was 1.9. Although SWB change scores were significantly correlated with GRC at month 1 and with PSR at months 1 and 3, the magnitudes of these correlations were between 0.22 and 0.26, which are considered small to moderate [8]. It is possible that using anchors that are weakly correlated with SWB underestimated the MID. Additional research is needed to identify a range of plausible MIDs for the SWB subscale using anchors that are clinically and subjectively meaningful to physicians and patients, as well as strongly correlated ($r > 0.3$) to SWB change scores.

The TOI was the primary HRQL end point of the IRIS trial. The lower end of the MID range for the TOI (i.e., 5 points) was prespecified for use in interpreting treatment group differences. It was concluded that TOI scores in the imatinib arm were higher (i.e., better physical function and well-being) than in the IFNα + LDAC arm, and that the difference was both statistically significant and clinically meaningful [39]. Furthermore, knowing the MID for the TOI allowed individual patients in the IRIS trial to be classified based on changes in their TOI scores. Hahn et al. classified patients into three categories: clinically relevant improvement (TOI increased 5 or more points from baseline), clinically relevant decline (TOI decreased 5 or more points from baseline), and no change. Clinically relevant declines in TOI were experienced by 52% to 73% of patients in the IFNα + LDAC arm compared with only 22% to 29% of patients in the imatinib arm. Clinically relevant improvements in TOI were experienced in 9% to 25% for IFNα + LDAC compared with 29% to 43% for imatinib [39]. HRQL results presented in this manner can easily be understood by both clinicians and patients. Another use of MIDs is for estimating sample size or power for a future study. For example, investigators can determine the number of subjects needed to detect an HRQL score difference equal to the MID or larger.

We used two types of anchors: the GRC, which was a retrospective, patient-reported assessment of change, and change in PSR, which was a prospective, physician-reported measure. One or both of these properties (i.e., retrospective vs. prospective, patient-reported vs. physician-reported) may be responsible for the poor agreement between GRC and PSR change (Table 3). Nevertheless, strong agreement between PSR change and GRC_{SWB} or GRC_{EWB} was not expected because PSR is a measure of change in physical functioning, whereas GRC_{SWB} and GRC_{EWB} are measures of change in social and emotional well-being respectively.

Using correlations between HRQL scores and anchors to assess the usefulness of anchors for an MID analysis has been suggested [7] and implemented [23]. We considered only those HRQL change scores that were statistically significantly related to the anchors. Had we failed to implement this criterion, the MIDs might have been underestimated. Because correlations between GRC and HRQL change scores were largest at the month 1 assessment and declined thereafter, the most suitable time to administer the GRC in this trial was at month 1, which measured perceived change from baseline. Our general recommendation is that a GRC should be administered at a time during a clinical trial when patients are expected to experience sufficient change in their health and HRQL. Thus, this finding is specific to this clinical trial and is not necessarily generalizable to other clinical trials. For example, in another clinical trial there may be a delayed response to therapy, and patients may experience little or no change in their HRQL from baseline to month 1, yet changes from month 5 to month 6 may be dramatic. In such an example, GRC administered at month 1 may not provide useful data for assessing clinically meaningful change.

Our study is not without limitations. Guyatt et al. [41] contend that if a transition rating, such as the GRC, were perfectly valid, it would have correlations with the pre- and post-HRQL scores of equal magnitude and opposite sign. We did not observe this pattern with the GRC data in this study. Correlations between GRC_{TOI} and TOI pre scores (i.e., baseline, month 2 or month 5) were small but positive and correlations between GRC_{TOI} and TOI post scores (i.e., months 1, 3, or 6) were larger and positive, indicating that patients' retrospective assessment of change may be overly influenced by their current health state, a potential problem with transition ratings that has been reported elsewhere [41]. Similar patterns for the pre- and post correlation coefficients were observed for the SWB and EWB subscales. These findings suggest that GRC may not have been the most sensitive measure of change at all assessment periods evaluated in this study.

We used a 5-point GRC scale, which we further collapsed into 3 categories of change. Others have used 7-point [16] or 15-point [15,41] GRC scales. Nevertheless, 7 or 15 categories of change may be too many to provide useful data for estimating the MID without collapsing categories. Sample sizes were lower than desired for some of the anchor-based assessments. We attempted to mitigate this problem by combining the "much worse" and "worse" categories and by combining the "much better" and "better" categories. This has a potential consequence of possibly increasing the magnitude of change scores in the combined categories, which could have led to overestimation of some MIDs. Nevertheless, our results for recommended MID ranges based on GRC or PSR change represented by five categories (data not shown) versus three categories were the same. Finally, our findings may be influenced by response shift over time [42] or recall bias associated with individual perceptions of change over the course of treatment.

We utilized an approach for identifying MIDs that combined distribution-based estimates and both patient- and physician-reported anchor-based estimates. We refined the anchor-based approach by using correlation coefficients to assess the appropriateness of the anchors to identify MID estimates for individual HRQL outcomes. We also improved on existing methods by ensuring the MID range included the next score larger than the variability of the scale and by using the interquartile range to identify plausible ranges of MIDs. While we illustrated this approach with the FACT-BRM in the context of a clinical trial, it can be easily and successfully applied to other HRQL instruments and research settings.

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

References

1 Cella DF. Measuring quality of life in palliative care. Semin Oncol 1995;22:73–81.

2 de Haes JC, Stiggelbout AM. Assessment of values, utilities and preferences in cancer patients. Cancer Treat Rev 1996;22(Suppl. A):S13–26.

3 Johnson JR, Temple R. Food and Drug Administration requirements for approval of new anticancer drugs. Cancer Treat Rep 1985;69:1155–9.

4 Clancy CM, Eisenberg JM. Outcomes research: measuring the end results of health care. Science 1998;282:245–6.

5 Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. Ann Intern Med 1993;118:622–9.

6 Lydick E, Epstein RS. Interpretation of quality of life changes. Qual Life Res 1993;2:221–6.

7 Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol 2003;56:395–407.

8 Cohen J. Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers, 1988.

9 Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. Control Clin Trials 1991;12(Suppl.):S142–58.

10 Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Med Care 1989;27(Suppl.):S178–89.

11 Wyrwich KW, Nienaber NA, Tierney WM, Wolinsky FD. Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life. Med Care 1999;37:469–78.

12 Wyrwich KW, Tierney WM, Wolinsky FD. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. J Clin Epidemiol 1999;52:861–73.

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

14 Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. J Consult Clin Psychol 1991;59:12–19.

15 Cella D, Hahn EA, Dineen K. Meaningful change in cancer-specific quality of life scores: differences between improvement and worsening. Qual Life Res 2002;11:207–21.

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 Cella D, Bullinger M, Scott C, Barofsky I. Group vs. individual approaches to understanding the clinical significance of differences or changes in quality of life. Mayo Clin Proc 2002;77:384–92.

18 Eton DT, Cella D, Yost KJ, et al. Minimally important differences on the functional assessment of cancer therapy-breast (FACT-B) scale: results from ECOG study 1193. J Clin Epidemiol 2004;57:898–910.

19 Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials 1989;10: 407–15.

20 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.

21 Wright JG. Interpreting health-related quality of life scores: the simple rule of seven may not be so simple. Med Care 2003;41:597–8.

22 Cella DF, Eton DT, Fairclough DL, et al. What is a clinically meaningful change on the Functional Assessment of Cancer Therapy-Lung (FACT-L): results from the Eastern Cooperative Oncology Group (ECOG) study 5592. J Clin Epidemiol 2002;55:285–95.

23 Yost K, Eton DT, Cella D, et al. Minimal important differences on the functional assessment of cancer therapy-coloretal. Qual Life Res 2002;11: 629.

24 Cella DF, Tulsky DS, Gray G, et al. The Functional Assessment of Cancer Therapy scale: development and validation of the general measure. J Clin Oncol 1993;11:570–9.

25 Cella DF. Quality of life outcomes: measurement and validation. Oncology (Huntingt) 1996;10: 233–46.

26 Cella D. Manual of the Functional Assessment of Chronic Illness Therapy (FACIT) Measurement System—Version 4. Evanston, IL: Center on Outcomes, Research and Education (CORE), Evanston Northwestern Healthcare and Northwestern University, 1997.

27 Bacik J, Mazumdar M, Murphy BA, et al. The Functional Assessment of Cancer Therapy-BRM (FACT-BRM): a new tool for the assessment of quality of life in patients treated with biologic response modifiers. Qual Life Res 2004;13:137–54.

28 Mazumdar M, Bacik J, Cella D, Fairclough D. Validation of FACT-BRM (Biological Response Modifier) and handling of missing data in QoL analysis. Society for Clinical Trials, 2000:112.

29 Motzer RJ, Murphy BA, Bacik J, et al. Phase III trial of interferon alfa-2a with or without 13-cis-retinoic acid for patients with advanced renal cell carcinoma. J Clin Oncol 2000;18:2972–80.

30 O'Brien SG, Guilhot F, Larson RA, et al. Imatinib compared with interferon and low-dose cytarabine for newly diagnosed chronic-phase chronic myeloid leukemia. N Engl J Med 2003;348:994–1004.

31 Beckerman H, Roebroeck ME, Lankhorst GJ, et al. Smallest real difference, a link between reproducibility and responsiveness. Qual Life Res 2001;10: 571–8.

32 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.

33 Guyatt GH, Osoba D, Wu AW, et al. Methods to explain the clinical significance of health status measures. Mayo Clin Proc 2002;77:371–83.

34 Farivar SS, Liu H, Hays RD. Another look at the half standard deviation estimate of the minimally important difference in health-related quality of life scores. Expert Rev Pharmacoeconomics Outcomes Res 2004;4:515–23.

35 Cella D, Eton DT, Lai JS, et al. 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 2002;24:547–61.

36 Ringash J, Bezjak A, O'Sullivary B, Redelmeier D. Interpreting differences in quality of life: the FACT-H&N in laryngeal cancer patients. Qual Life Res 2004;13:725–33.

37 Wells GA, Tugwell P, Kraag GR, et al. Minimum important difference between patients with rheumatoid arthritis: the patient's perspective. J Rheumatol 1993;20:557–60.

38 Wyrwich KW, Fihn SD, Tierney WM, et al. Clinically important change in health-related quality of life for patients with chronic obstructive pulmonary disease. J Gen Intern Med 2003;18:196–202.

39 Hahn EA, Glendenning GA, Sorensen MV, et al. Quality of life in patients with newly diagnosed

chronic phase chronic myeloid leukemia on imatinib versus interferon alfa plus low-dose cytarabine: results from the IRIS Study. J Clin Oncol 2003;21:2138–46.

40 Cella D, Yost KJ, Lai JS, Zagari MA. General US population norms for the Functional Assessment of Cancer Therapy-General (FACT-G). Qual Life Res 2003;12:852.

41 Guyatt GH, Norman GR, Juniper EF, Griffith LE. A critical look at transition ratings. J Clin Epidemiol 2002;55:900–8.

42 Schwartz CE, Sprangers MA. Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. Soc Sci Med 1999;48:1531–48.