VALUE IN HEALTH 18 (2015) 846–855

Available online at www.sciencedirect.com

ScienceDirect

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

Deriving a Preference-Based Measure for Myelofibrosis from the EORTC QLQ-C30 and the MF-SAF

Clara Mukuria, PhD[1,*], *Donna Rowen, PhD*[1], *John E. Brazier, PhD*[1], *Tracey A. Young, PhD*[1], *Beenish Nafees, MSc*[2]

[1]*Health Economics and Decision Science, School of Health and Related Research, The University of Sheffield, Sheffield, UK;* [2]*ICON plc, Dublin, Ireland*

ABSTRACT

Background: Utility values are required for economic evaluation using cost-utility analyses. Often, generic measures such as the EuroQol five-dimensional questionnaire are used, but this may not appropriately reflect the health-related quality of life of patients with cancer including myelofibrosis. **Objective:** To derive a condition-specific preference-based measure for myelofibrosis using appropriate existing measures, the Myelofibrosis-Symptom Assessment Form and the European Organisation for Research and Treatment of Cancer Quality of Life 30 Questionnaire. **Methods:** Data from the Controlled Myelofibrosis Study with Oral JAK Inhibitor Treatment trial (n = 309) were used to derive the health state classification system. Psychometric and factor analyses were used to determine the dimensions of the classification system. Psychometric and Rasch analyses were then used to select an item to represent each dimension. Item selection was validated with experts. A selection of health states was valued by members of the general population using time trade-off. Finally, health state values were modeled using regression analysis to produce utility values for every state. **Results:** The Myelofibrosis 8 dimensions has eight dimensions: physical functioning, emotional functioning, fatigue, itchiness, pain under ribs on the left side, abdominal discomfort, bone or muscle pain, and night sweats. Regression models were estimated using time trade-off data from 246 members of the general population valuing a total of 33 states. The best performing model was a random effects maximum likelihood model producing utility values ranging from 0.089 to 1. **Conclusions:** The Myelofibrosis 8 dimensions is a condition-specific preference-based measure for myelofibrosis. This measure can be used to generate utility values for myelofibrosis for any data set containing the Myelofibrosis-Symptom Assessment Form and the European Organisation for Research and Treatment of Cancer Quality of Life 30 Questionnaire data.

Keywords: EORTC QLQ-C30, MF-SAF, myelofibrosis, preference-based measure.

Introduction

Preference-based measures of health are used in cost-utility analysis to inform health policy, such as by agencies like the National Institute for Health and Care Excellence in England [1]. They are used to generate utility values for health-related quality of life (HRQOL), which are combined with length of life to generate a quality-adjusted life-year. The current National Institute for Health and Care Excellence reference case [1] prefers utility values from the EuroQol five-dimensional questionnaire (EQ-5D) [2]. The EQ-5D, however, may not be appropriate for all patient groups or all populations [3,4].

Myelofibrosis (MF) is a cancer characterized by scarring of the bone marrow, progressive anemia, and enlarged spleen. Symptoms include consequences of enlarged spleen (pain or fullness below the ribs on the left side, feeling full sooner than normal when eating) and constitutional symptoms such as fatigue, itching, weight loss, dyspnea, fever, and night sweats [5]. The EQ-5D has been used in some cancer populations in which it has been shown to be valid [4]. It has been argued, however, that the EQ-5D does not appropriately reflect the HRQOL of all patients with cancer [6].

A condition-specific measure has been successfully used in the MF population, the Myelofibrosis-Symptom Assessment Form version 2.0 (MF-SAF 2.0). The MF-SAF 2.0 is based on the MF-SAF [5,7], with seven items focusing on specific MF symptoms: abdominal discomfort, pain under left ribs, early satiety, night sweats, itching, bone or muscle pain, and inactivity. It was developed to assess the specific symptoms of MF and other myeloproliferative neoplasm conditions such as polycythemia vera or essential thrombocythemia. The more generic cancer measure the European Organisation for Research and Treatment of Cancer Quality of Life 30 Questionnaire (EORTC QLQ-C30) has also been used in this population. The EORTC QLQ-C30 is one of the most commonly used generic cancer measures [8] consisting of 30 questions across six functioning scales (physical, role,

* *Address correspondence to:* Clara Mukuria, Health Economics and Decision Science, School of Health and Related Research, University of Sheffield, Regent Court, 30 Regent Street, Sheffield S1 4DA, UK.

E-mail: c.mukuria@sheffield.ac.uk.

1098-3015$36.00 – see front matter © 2015 Published by Elsevier Inc. on behalf of International Society for Pharmacoeconomics and Outcomes Research (ISPOR).

http://dx.doi.org/10.1016/j.jval.2015.07.004

cognitive, emotional, social, and global quality of life) and nine symptom scales (fatigue, nausea and vomiting, pain, dyspnea, sleep disturbance, appetite loss, constipation, diarrhea, and financial impact). Although the EORTC QLQ-C30 has been designed to be applicable for assessing generic aspects of quality of life, it has limitations and a modular approach has been taken in which cancer-specific modules are added to the core EORTC QLQ-C30 [9]. Assessment of the psychometric performance of the QLQ-C30 against MF measures indicate that the QLQ-C30 captures functioning and some generic symptom problems [10,11]. However, it does not cover MF-specific symptoms, which may be important in this population, such as weight loss, itching, and night sweats, nor is it as responsive as the MF-SAF 2.0 over time [12]. There was also evidence that dimensions such as constipation and diarrhea were less relevant for MF populations [12]. A number of validated EORTC QLQ-C30 add-on cancer-specific modules exist, but there is none for MF.

A further limitation of these measures is that they are not preference-based, though a preference-based measure has been derived from the EORTC QLQ-C30, the EORTC-8D [13], which allows utility values to be estimated. This algorithm as well as mapping functions between the QLQ-C30 and the EQ-5D has been estimated in a population with MF [14]. Given the lack of overlap with some MF-related symptoms noted above, neither the preference-based EORTC-8D nor mapping was considered sufficient for providing utility values for MF populations. The aim of this study was therefore to derive a preference-based measure for MF from the MF-SAF and the QLQ-C30 that captures the HRQOL of patients with MF.

Methods

This study involved a five-step process to derive a health state classification system using an existing data set followed by valuation of the classification system to generate utility values based on methods originally developed for this purpose by Brazier et al. [15] and subsequently to generate the EORTC-8D [13].

Data

Data were from the Controlled Myelofibrosis Study with Oral JAK Inhibitor Treatment (COMFORT-I) trial, a randomized, double-blind, placebo-controlled phase III study of the oral JAK1/JAK2 inhibitor INCB018424 (ruxolitinib) in patients with primary MF, post–polycythemia vera MF, or post–essential thrombocythemia MF conducted in 68 sites in the United States, 6 sites in Canada, and 15 in Austria [16]. A total of 309 participants were recruited, and 248 completed the study. Their mean age was 67 ± 8.78 years (range, 40–91 years), and 54% were men. Respondents completed a number of questionnaires including both the MF-SAF 2.0 and the EORTC QLQ-C30.

MF-SAF 2.0 seven-item scores range from 0 (“absent” symptoms) to 10 (“worst imaginable” symptoms), and the total symptom score is the sum of the individual scores, excluding inactivity (see Appendix 1 in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2015.07.004). In COMFORT-I, the MF-SAF 2.0 diary was collected daily over 24 weeks.

The EORTC QLQ-C30 questions [8] from five of the functioning scales and nine of the symptom scales were used in the analysis. The global quality-of-life and financial impact items are unsuitable for inclusion in a health state classification system of a preference-based measure and are excluded from the analyses. The QLQ-C30 was administered at baseline and 24-week follow-up.

Analysis

The aim was to produce a classification system amenable to valuation by respondents with a minimum loss of information while ensuring that the MF-SAF and QLQ-C30 responses could be unambiguously converted into the classification system. The analysis followed five steps [15] that have been used to derive other condition-specific preference-based measures from both single measures [17,18] and a battery of measures used in epilepsy [19]. This study extends the latter approach by deriving a single condition-specific preference-based measure from two measures.

Step 1: Dimensional Structure

The dimensionality of the QLQ-C30 and the MF-SAF 2.0 was assessed separately for each measure and together for both measures using factor analysis to determine the dimensions across both measures and all items. Separate assessment allowed the identification of different factors within each measure, whereas joint assessment allowed the identification of common factors across the measures. Scree plots were used to confirm the number of factors with eigenvalues greater than 1 used as a cutoff for inclusion of factors. Item loading was used to assess which factor the item was contributing toward, with higher absolute values indicating greater contribution [20]. Variance of the item not explained by the common factors and uniqueness were used to indicate items that may be measuring something outside the common factors.

Step 2: Item Selection

In the preference-based measure, each dimension needs to be represented using the minimum number of items from the original measure and the selected item(s) must best represent the overall dimension. This selection was undertaken using classical psychometric analysis and Rasch analysis.

Conventional psychometric tests were used to assess the practicality and validity of the items. A strong item for selection should have low levels of missing data, high correlation with the dimension score, and responses across the severity range and should be able to discriminate between severity levels and be responsive to change over two points in time [21,22].

Rasch analysis is a mathematical technique that converts categorical item responses to a continuous latent scale using a logit model [23,24]. It was used to assess whether items fitted the dimension and that they covered the severity of the health problem as well as the extent to which items had response choices that were appropriately ordered and whether items perform the same between populations (testing for differential item functioning) [21,22]. Items that did not meet these criteria were candidates for being excluded from the classification system. An equality test and selection of divergent items using a t test were undertaken to confirm the dimension structure of the MF-SAF. Local dependency was examined to determine whether items were redundant. Rasch analysis was undertaken on the MF-SAF using the average data for days 1 to 7 (baseline) excluding the inactivity item. Validation was done using MF-SAF day 1 data. Results from previous Rasch analyses on the properties of the QLQ-C30 items were used to select items from the QLQ-C30 [13].

Step 3: Validation of the Classification System

Face validity of the MF-8D classification system was examined using a small sample of clinicians who had patients with MF. They were asked their opinions on each of the proposed dimensions including whether the aspect of HRQOL was valid as a

separate category, and also asked to consider whether additional or alternative dimensions would improve the classification system.

Step 4: Valuation

It is infeasible to value all health states defined by the classification system because the total number is too large. A sample of health states was selected for valuation using an orthogonal array that enables the estimation of an additive regression model estimating utility values for every health state. An orthogonal array generated using IBM SPSS statistics version 21 [25] selected 32 health states. These were first divided into mild and moderate states and then randomly allocated in turn into four combinations of eight health states, known as "blocs." Each respondent valued health states in one card bloc, plus the worst health state defined by the classification system.

Valuation survey

Selected health states were valued by a general population sample using time trade-off (TTO). Interviewers were each provided a sex and age quota of respondents to ensure the interviewee sample was representative of the UK general population in terms of age and sex. Each interviewer found willing participants using convenience methods.

At the interview, respondents were provided with a description of MF (see Appendix 2 in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2015.07.004). Respondents were asked to read and complete the MF-8D classification system for their own health to familiarize themselves with the classification system. Respondents then valued nine health states (one card bloc plus the worst health state) plus "full health" and "dead" using the visual analogue scale, where 100 represents "best possible state of health" and 0 represents "worst possible state of health." This was a warm-up task to encourage respondents to think about the ordering of the health states and their relative value.

This was followed by a practice TTO question using a moderate health state and then by a TTO task for the nine health states presented in a random order. Respondents were first asked whether the health state was better or worse than being dead. For health states better than dead, the Measurement and Valuation of Health TTO protocol with a visual prop design was used [2]. Respondents were asked to choose between 1) health state h for 10 years, after which they will die, or 2) full health for z years ($z \leq 10$), after which they will die. The ping-pong titration method was used to determine the value of z when respondents are indifferent between 1) and 2), when the number of years in full health changes from 10 to 0.5, from 9.5 to 1, and so forth. For health states worse than dead, the lead-time TTO method was used [26]. Respondents were asked to choose between 1) 10 years in full health plus health state h for 10 years, after which they will die, or 2) full health for z years ($z \leq 10$), after which they will die. The ping-pong method was used to determine the value of z when respondents are indifferent between 1) and 2). At the end of the interview, respondents self-completed the EQ-5D and sociodemographic questions.

Valuation data

There were 252 successfully conducted interviews. A total of 244 respondents are included in the analyses. Eight respondents were excluded from the analysis; four for valuing all states as identical and less than one; two for valuing PITS, that is, the worst health state, higher than every other state; and two for valuing all health states as worse than being dead because arguably they did not think life was worth living in any impaired state. Table 1 summarizes the characteristics of these respondents and compares these characteristics with characteristics of the general population of England. The valuation sample contained a higher proportion of people aged 18 to 40 years, more women, more employed people, more full-time students, fewer retired people, and fewer people in poor health.

Table 1 – Characteristics of respondents in the valuation survey.

Characteristic	Included respondents (n = 244)*	England†
Age (y), mean ± SD	38.0 ± 14.8	38.6
Age distribution (y) (%)		
18–40	61.1	41.6
41–65	33.6	39.1
>65	5.3	19.3
Sex: female (%)	62.3	51.3
White British (%)	82.0	
Employed or self-employed (%)	69.7	60.9
Unemployed (%)	3.7	3.4
Full-time student (%)	16.8	7.3
Retired (%)	6.1	13.5
Secondary school is highest level of education (%)	11.9	NA
Completed university (%)	59.0	
EQ-5D-5L score, mean ± SD	0.92 ± 0.13	0.86 ± 0.23†

EQ-5D-5L, five-level EuroQol five-dimensional questionnaire; NA, not applicable.

* Eight respondents were excluded: four for valuing all states as identical and less than 1; two for valuing PITS higher than every other state; and two for valuing all states as worse than being dead.

† Interviews conducted in the Measurement and Valuation of Health (MVH) study in 1993 [27].

Step 5: Modeling the Utility Values

Utility values were modeled using regression analysis to estimate values for every health state defined by the classification system. The specification is as follows:

$$y_{ij} = f(\mathbf{X}_{\delta\lambda}\boldsymbol{\beta}) + \varepsilon^{s}_{ij} \quad (1)$$

The dependent variable y_{ij} is TTO disvalue (generated as 1 – TTO) for health state i valued by respondent j, and $\mathbf{X}_{\delta\lambda}$ is a vector of dummy explanatory variables for each level λ of dimension δ of the classification system. Level $\lambda = 1$ acts as a baseline for each dimension, and ε_{ij} is the error term. For dimensions in the classification system derived from the MF-SAF questionnaire, only one dummy variable is included when the dimension is at the severity level "worst imaginable." The error term $\varepsilon_{ij} = u_j + e_{ij}$ can be divided into u_j the individual random effect and e_{ij} the random error term for the ith health state valuation of the jth individual.

Models were estimated using ordinary least squares with robust standard errors, and random effects models with robust standard errors were also estimated to take into account differences at the individual level because these models appropriately deal with the structure of the data when each respondent has multiple observations [15]. Tobit models that take into account the bounded nature of the data that is bounded at 1 and –1 were estimated. Inconsistent coefficients for adjacent severity levels of a dimension indicate that health deterioration leads to a higher utility value, contrary to expectations. Models that merge

inconsistent adjacent severity levels to remove these inconsistencies were estimated.

Performance of regression models was assessed using the number of inconsistent coefficients and significant coefficients' mean absolute error (MAE) of predictions at the health state level and the number of health states in the valuation survey where MAE is greater than 5% and 10%. MAE was generated using the difference between observed and predicted utility values at the health state level, and models with a lower MAE were preferred.

Ethical review for the valuation and content validity studies was obtained from the Salus Institutional Review Board (Study no. 0050-0250).

Results

Step 1: Dimensional Structure

Factor analysis indicated that the combined QLQ-C30 and MF-SAF items loaded onto seven factors (see Appendix 3 in Supplemental Materials found at 10.1016/j.jval.2015.07.004). There was a single physical, role, and social functioning, fatigue, and dyspnea factor with the inactivity item from the MF-SAF. The remaining MF-SAF items loaded onto the second factor with the pain items from the QLQ-C30. Early satiety was strongly correlated with pain, but itching and night sweats may have loaded onto this otherwise pain-related factor because they were highly associated with MF rather than because they belong to the same factor. Emotional and cognitive functioning were separate factors, and the remaining three factors had items relating to digestion and sleep disturbance. Factor analysis of the QLQ-C30 and the MF-SAF separately confirmed these results (available on request). Further assessment in steps 2 and 3 was required to confirm the dimensions.

Step 2: Item Selection

The psychometric analysis indicated that all the MF-SAF items performed relatively well on the basis of the analysis of missing data, floor effects, internal consistency, and responsiveness although there was some evidence of ceiling effects (Table 2). For the QLQ-C30, items related to physical functioning (pf1 and pf2), emotional functioning (ef2 and ef3), all fatigue items, dyspnea, sleep disturbance, and appetite loss performed well too. Items with very high ceiling effects such as pf5, nausea, vomiting, and constipation suggest that these were not common symptoms for this population. All QLQ-C30 items for role, cognitive, and social functioning and symptoms related to eating and digestion (vomiting, pain, constipation, and diarrhea) had low to very small standardized response means (SRMs) ($<|0.20|$), suggesting that these items were not suitable for measuring change. Items with high levels of ceiling or floor effects and/or low responsiveness were excluded from the classification system.

The threshold maps from the Rasch analysis show that the six MF-SAF items were correctly ordered across the 11 severity levels (Fig. 1). Inactivity was excluded from the Rasch analysis on the basis that it belonged to a different factor (see step 1). Nineteen respondents had extreme values (person fit residuals $\geq|4|$), and these were excluded from further analysis because their scores can affect the logit scaling. The Rasch analysis on the MF-SAF indicated that items 3 (abdominal discomfort) and 5 (early satiety) suffered from local dependency and were tapping into the same aspect of HRQOL, meaning that one of the items is redundant and should be excluded (Table 3). Item 6 had uniform sex differential item functioning; therefore, it was also a candidate for exclusion because this item performed differently for men and women. Rasch models were reestimated excluding items 3 or 5 and/or item 6, finding that excluding either item 3 or 5 improved model fit and excluding item 6 along with either item 3 or 5 also improved model fit.

MF-SAF items 1, 2, 3, and 4 were selected, covering night sweats, itching, abdominal discomfort, and rib pain. Item 3 (abdominal discomfort) was selected instead of item 5 (early satiety) because of potential problems with valuation of "feelings of fullness" (see step 3). The QLQ-C30 items for physical functioning (2-pf2 and 3-pf3), fatigue (12-fa3), and dyspnea (8-dy1) were selected, and the wording from the EORTC-8D was used for physical functioning and fatigue. Item 22 (ef2 worry) was included for the emotional functioning dimension because this had the highest SRM for this dimension. Sleep disturbance and appetite loss were excluded because the former was associated with physical functioning and fatigue while the latter was associated with early satiety.

The proposed classification system for validation in step 3 had eight dimensions: physical functioning, emotional functioning, fatigue, and dyspnea (from the QLQ-C30) and itching, rib pain on the left side, abdominal discomfort, and night sweats (from the MF-SAF). The MF-SAF items had two severity levels, "absent" and "worst imaginable," in the classification system for valuation because it is infeasible to value 11 severity levels as members of the general population are unlikely to be able to distinguish between these levels.

Step 3: Validation of the Classification System

Four MF specialists were asked whether the proposed dimensions were valid and separate aspects of HRQOL. They were also asked to consider whether an additional dimension of "bone or muscle pain" should be included, and whether "feeling of fullness" could be included as an alternative dimension to "abdominal discomfort." The specialists noted that "shortness of breath" was important only when doing activities, but the QLQ-C30 question did not state this. They also felt that there was overlap between this item and physical functioning; therefore, this was excluded from the classification system. Abdominal discomfort rather than feeling of fullness was included. All the specialists felt that bone or muscle pain was an important and distinct symptom of MF that should be included, and this was therefore added to the classification system. The final selected classification system had eight dimensions: physical functioning, emotional functioning, and fatigue (from the QLQ-C30) and itching, rib pain on the left side, abdominal discomfort, bone or muscle pain, and night sweats (from the MF-SAF) (Fig. 2). The classification described a total of 2560 health states.

Step 4: Valuation Survey

Descriptive statistics for observed TTO utility values for all health states included in the valuation survey (ordered by mean health state utility value) show that in general the misery score (generated using the summed severity levels of each health state using equal severity weighting for each dimension) decreases as the mean and median utility values decrease (see Appendix 3 in Supplemental Materials found at 10.1016/j.jval.2015.07.004). The precise ordering of the misery score and the mean TTO value, however, differ. This may occur because the misery score assumes that all dimensions are equally weighted whereas this may not be true for utility values. In addition, the difference in mean utility values between many health states is small, at 0.01 or 0.02. Utility values were distributed across the full potential range from –1 to 1. Most of the values were greater than zero, meaning that the health state was regarded as being better than dead, but there was also a peak of values at –1, with few values lying between –1 and 0.

Table 2 – Psychometric analysis of MF-SAF and QLQ-C30 items (n = 263).

Item no.	Item wording	% missing data	% response at floor (very much/worst)	% response at ceiling (not at all/absent)	Correlation*,†	SRM
MF-SAF						
MF-SAF01	Worst night sweats (flushed)	3.6	1.1	16.9	0.724	–0.31
MF-SAF02	Worst itchiness	3.6	0.8	25.6	0.699	–0.33
MF-SAF03	Worst abdominal discomfort	3.6	1.1	5.3	0.901	–0.36
MF-SAF04	Worst pain under ribs on left	3.6	0.8	18.4	0.842	–0.39
MF-SAF05	Worst feeling fullness (early satiety)	3.6	0.4	7.1	0.872	–0.34
MF-SAF06	Worst bone or muscle pain	3.6	1.1	18.0	0.756	–0.20
MF-SAF07	Worst degree of inactivity	3.6	0.4	7.5	Not in total	–0.29
	EORTC QLQ-C30				**Domain score**	
	Physical functioning					
pf1	Trouble with strenuous activities?	4.2	17.7	14.3	–0.833	–0.19
pf2	Trouble with taking a long walk?	3.9	24.1	12.8	–0.893	–0.21
pf3	Trouble taking a short walk?	3.9	2.6	52.6	–0.824	–0.08
pf4	Stay in bed or chair during the day?	3.9	3.0	45.5	–0.671	–0.15
pf5	Help eating, dressing, washing?	4.5	0.4	94.4	–0.335	0.05
	Role functioning					
rf1	Limited in work or daily activities?	4.2	8.6	32.3	–0.910	0.07
rf2	Limited in pursuing hobbies?	4.5	11.3	33.8	–0.919	0.00
	Emotional functioning					
ef1	Did you feel tense?	4.9	3.8	47.4	–0.786	–0.10
ef2	Did you worry?	4.2	4.5	33.8	–0.866	–0.24
ef3	Did you feel irritable?	4.2	3.4	41.0	–0.814	–0.20
ef4	Did you feel depressed?	5.2	3.0	50.8	–0.811	–0.06
	Cognitive functioning					
cf1	Difficulty concentrating on things?	4.2	0.8	59.8	–0.799	–0.05
cf2	Have had difficulty remembering things?	4.9	1.9	49.2	–0.830	–0.02
	Social functioning					
sf1	Interfered with family life?	4.2	6.0	40.6	–0.910	–0.04
sf2	Interfered with social activities?	4.9	9.0	28.6	–0.928	–0.04
	Fatigue					
fa1	Did you need to rest?	4.2	13.9	8.3	0.873	–0.23
fa2	Have you felt weak?	4.2	13.9	15.4	0.865	–0.23
fa3	Were you tired?	4.9	21.1	5.3	0.886	–0.26
	Nausea and vomiting					
ns1	Have you felt nauseated?	4.5	1.1	67.7	0.983	–0.25
vt1	Have you vomited?	4.2	0.0	92.9	0.504	–0.03
	Pain					
pa1	Have you had pain?	4.5	5.3	31.6	0.931	0.00
pa2	Pain interferes with daily activities?	4.5	4.1	49.6	0.917	–0.02
dy1	Were you short of breath?	4.5	9.0	26.7	–	–0.20
sl1	Have you had trouble sleeping?	4.9	10.5	26.3	–	–0.25
al1	Have you lacked appetite?	4.2	7.5	36.8	–	–0.33
co1	Have you been constipated?	3.9	2.6	67.3	–	–0.07
di1	Have you had diarrhea?	4.5	2.3	56.8	–	–0.03

MF-SAF, Myelofibrosis-Symptom Assessment Form; QLQ-C30, Cancer Quality of Life 30 Questionnaire; SRM, standardized response mean.
* Values for correlation for the MF-SAF are for MF-SAF total.
† Values for correlation for the EORTC QLQ-C30 are for domain score.

Step 5: Modeling Health State Utility Values

The ordinary least squares, random effects maximum likelihood estimator (RE MLE), and Tobit models had inconsistent but non-significant coefficients for physical functioning level 2 and fatigue level 2 (Table 4). In addition, the ordinary least squares model had an inconsistent coefficient for fatigue level 3 and the RE MLE model had an inconsistent coefficient for physical functioning level 4. This means that contrary to expectations as health worsens utility increases.

Overall, the RE MLE performs best taking into account the number of significant level coefficients, inconsistencies, and the MAE. Model 4 is a consistent version of the RE MLE model (2), in which levels of inconsistent coefficients were merged with the

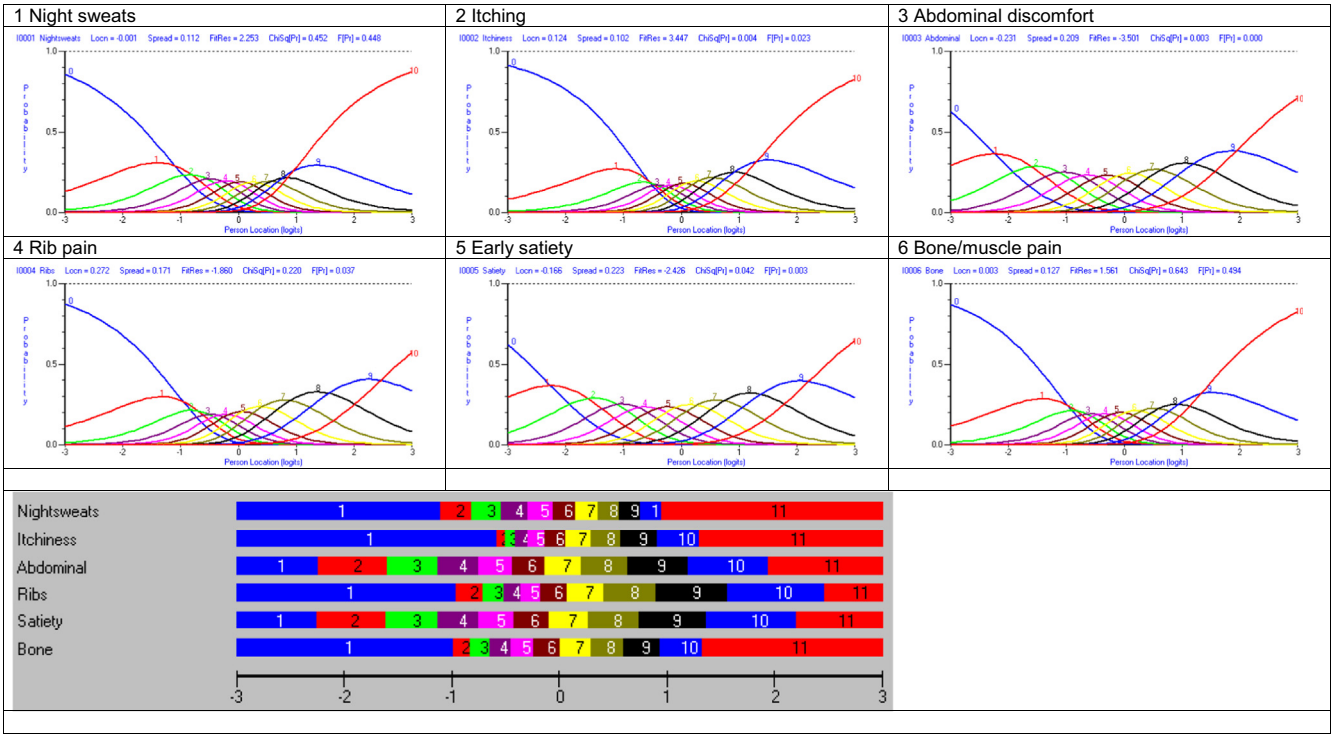

Fig. 1 – MF-SAF item probability curves and threshold map using average data. MF-SAF, Myelofibrosis-Symptom Assessment Form. (Color version of figure is available online.)

Table 3 – Myelofibrosis-Symptom Assessment Form (MF-SAF) Rasch results (n = 302).

Item no.	Item wording	Item-level ordering	Local dependency	DIF (age)	DIF (sex)	DIF (MF type)	Fit residual	Item-level χ^2	Consider for exclusion
MF-SAF01	Night sweats (flushed)	Ordered					2.176	0.458	No
MF-SAF02	Itchiness	Ordered					3.352	0.005	No
MF-SAF03	Abdominal discomfort	Ordered	√(MFSAF05)				−3.425	0.003	Yes or MFSAF05
MF-SAF04	Pain under ribs on left	Ordered					−1.82	0.222	No
MF-SAF05	Feeling fullness (early satiety)	Ordered	√(MFSAF03)				−2.411	0.041	Yes or MFSAF03
MF-SAF06	Bone or muscle pain	Ordered			√		1.515	0.650	Yes
Criteria at the model level									
χ^2 interaction		<0.001							
Item fit		0 (0.18)							
Person fit		−0.79 (0.79)							
Person separation index		0.83							
Unidimensional (equality test—proportion with significant differences) (%)		7.8							

Note. DIF was tested for age, sex, and the type of MF reported. DIF was assessed on the basis of a Bonferroni-adjusted *P* value = 0.002778. Item-level fit assessed on the basis of a Bonferroni-adjusted *P* value = 0.0017.
√, presence of local dependency or DIF; DIF, differential item functioning; MF, myelofibrosis.

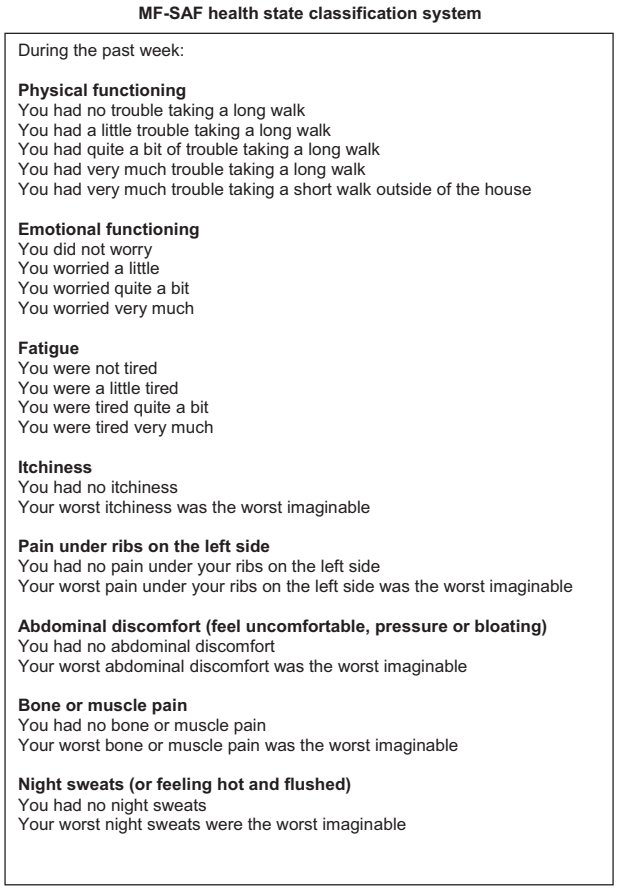

Fig. 2 – MF-SAF health state classification system. MF-SAF, Myelofibrosis-Symptom Assessment Form.

adjacent severity level to ensure that as health worsens the utility value decreases.

These regression models enable a utility value to be determined for every health state defined by the MF-8D classification system. For dimensions derived from the MF-SAF, however, there were only two severity levels in the classification system, absent and worst imaginable, whereas respondents were to respond on a 0 to 10 scale, where 0 = absent and 10 = worst imaginable. Table 5 reports the utility decrement for these items assuming equal intervals for each decrement on the 0 to 10 scale (method 1), and an alternative method using the relative difference in the logit value generated using the Rasch logit model for each decrement on the 0 to 10 scale for each individual dimension (method 2).

The MF-8D algorithm was applied to COMFORT-I data, which contained both the MF-SAF and the EORTC QLQ-C30. The MF-8D has utility scores ranging from 0.089 to 1. Mean observed utility scores in the data set were 0.732 ± 0.164 (range 0.226–1) at baseline and 0.785 ± 0.174 (range 0.110–1) using method 1. Mean was 0.669 ± 0.161 (range 0.230–1) at baseline and 0.726 ± 0.178 (range 0.104–1) at follow-up using method 2. The difference in utility scores for the two versions was statistically significant at baseline (mean difference 0.0630 ± 0.040; $t_{232} = 24.3$; $P < 0.0001$) and follow-up (mean difference 0.0589 ± 0.046; $t_{209} = 18.4$; $P < 0.0001$). As expected, both versions were able to discriminate between MF-SAF total symptom score groups (baseline MF-8D vs. method 1 $F_{231,3} = 271$, $P < 0.001$; MF-8D vs. method 2 $F_{231,3} = 203$; $P < 0.001$) and Eastern Cooperative Oncology Group [28] groups (baseline MF-8D vs. method 1 $F_{224,2} = 7.6$, $P < 0.001$; MF-8D vs. method 2 $F_{224,2} = 5.9$, $P = 0.003$). SRMs are 0.36 and 0.39 for utility scores generated using method 1 and method 2, respectively, which mirrors SRMs from the MF-SAF 2.0 items.

Discussion

This study derived the classification system and corresponding utility values for MF-8D, a condition-specific preference-based measure for MF derived from the MF-SAF and the generic cancer measure EORTC QLQ-C30. The generated MF-8D utility values can be used to estimate quality-adjusted life-years for use in economic evaluation using data from past and future myeloproliferative neoplasm conditions trials that have used both measures. The classification system has eight dimensions: physical functioning, emotional functioning, fatigue, night sweats, itchiness, abdominal discomfort, pain under the left rib, and bone or muscle pain, with four or five severity levels for the first three dimensions and two levels for the last five dimensions. The inclusion of MF-specific symptoms and the exclusion of symptoms such as nausea and vomiting that were not relevant in this population means that this measure is better suited for generating utility values than either the EORTC-8D or mapped values to the EQ-5D.

The RE MLE model was the best performing model, and the RE MLE model with consistent coefficients is recommended for use in generating utility values. Two methods were used to derive utility values for patient responses for the MF-8D dimensions derived from the MF-SAF: method 1 assumed equal intervals across the 0 to 10 scale, and method 2 used results from the Rasch logit model. Utility scores derived using both methods were able to discriminate across groups with known differences as well as capture change over time in a similar manner to the original measures. Method 2 has the advantage that it divides the utility weight in a way that is consistent with the impact of each decrement on the underlying health of the patient. There is no reason, however, to expect that this would equal the preferences of the general population regarding each decrement on the 0 to 10 scale, and therefore method 1 is the recommended method for generating utility values.

There are a number of limitations in this study. The MF-8D was derived from two measures, meaning that the MF-8D can be applied only when both measures have been used. A potential solution would be to provide a mapping algorithm between the MF-SAF 2.0 questions and the MF-8D that would allow utility values to be generated when only the MF-SAF 2.0 has been used (available on request).

The use of two measures to derive a single preference-based measure was a novel approach that has been applied once before where a battery of measures was used [19]. The rationale was to take advantage of important core cancer dimensions from the QLQ-C30 while ensuring that MF-specific dimensions were covered. This was supported by the psychometric analysis, which confirmed that many of the QLQ-C30 items were not relevant or as responsive in this population as the MF-SAF items. Factor analysis was undertaken separately for each measure before combining the measures to ensure that factors were not contaminated by instrument-specific effects. Rasch analysis was used in the MF-SAF to test the individual items for the single factor in this measure and to inform item selection [15].

Development of the classification system was undertaken using a single data set that did not cover the full severity range, but it has not been validated in an external data set because of lack of data. Clinicians provided input, however, on the face validity of the items, indicating that they were appropriate for the condition. There are also common dimensions of physical functioning, emotional functioning, and fatigue in both the MF-8D and the EORTC-8D, the preference-based measure derived from

Table 4 – Estimated preference weights for the MF-8D

MF-8D variable levels	1: Ordinary least squares	2: RE MLE	3: Tobit	MF-8D variables -collapsed levels	4: RE MLE consistent model
MF-8D Physical Functioning 2	–0.017 (0.416)	–0.014 (0.234)	–0.007 (0.564)	MF-8D Physical Functioning 2	0
MF-8D Physical Functioning 3	0.070* (0.003)	0.068* (0.000)	0.057* (0.000)	MF-8D Physical Functioning 3 and 4	0.074* (0.000)
MF-8D Physical Functioning 4	0.091* (0.002)	0.065* (0.000)	0.062* (0.001)		
MF-8D Physical Functioning 5	0.120* (0.000)	0.115* (0.000)	0.103* (0.000)	MF8D Physical Functioning 5	0.122* (0.000)
MF-8D Emotional Functioning 2	0.017 (0.440)	0.031† (0.023)	0.021‡ (0.085)	MF-8D Emotional Functioning 2	0.031† (0.021)
MF-8D Emotional Functioning 3	0.041‡ (0.054)	0.049* (0.002)	0.043* (0.000)	MF-8D Emotional Functioning 3	0.048* (0.002)
MF-8D Emotional Functioning 4	0.068* (0.002)	0.074* (0.000)	0.058* (0.000)	MF-8D Emotional Functioning 4	0.075* (0.000)
MF-8D Fatigue 2	–0.013 (0.560)	–0.012 (0.455)	–0.010 (0.402)	MF-8D Fatigue 2	
MF-8D Fatigue 3	–0.004 (0.844)	0.006 (0.704)	0.008 (0.541)	MF-8D Fatigue 3	0.013 (0.364)
MF-8D Fatigue 4	0.053† (0.016)	0.066* (0.000)	0.051* (0.000)	MF-8D Fatigue 4	0.072* (0.000)
MF-8D Itchiness 2	0.097* (0.000)	0.093* (0.000)	0.084* (0.000)	MF-8D Itchiness 2	0.093* (0.000)
MF-8D Pain under Left Ribs 2	0.145* (0.000)	0.139* (0.000)	0.123* (0.000)	MF-8D Pain under Left Ribs 2	0.139* (0.000)
MF-8D Abdominal Discomfort 2	0.142* (0.000)	0.145* (0.000)	0.127* (0.000)	MF-8D Abdominal Discomfort 2	0.145* (0.000)
MF-8D Bone/Muscle Pain 2	0.179* (0.000)	0.178* (0.000)	0.151* (0.000)	MF-8D Bone/Muscle Pain 2	0.178* (0.000)
MF-8D Night Sweats 2	0.073* (0.000)	0.080* (0.000)	0.065* (0.000)	MF-8D Night Sweats 2	0.080* (0.000)
Constant	0.029 (0.229)	0.020 (0.342)	0.050† (0.014)	Constant	0.007 (0.715)
Observations	2196	2196	2196		2196
R^2	0.251				
Number of ID		244	244		244
Inconsistencies	3	3	2		0
Significant-level coefficients	11	12	12		11
MAE	0.018	0.021	0.049		0.022
MAE > 0.05	3	3	8		3
MAE > 0.10	0	0	4		0

Note. *P* values in parentheses. MAE, mean absolute error; MF-8D, Myelofibrosis 8 dimensions; RE MLE, random effects maximum likelihood estimator.

* Significant at 1%.

† Significant at 5%.

‡ Significant at 10%.

the QLQ-C30, which provides some external validity [13]. Interestingly, the remaining five dimensions in each measure differ, suggesting that the dimensions of the HRQOL differ across different cancer populations, and this raises questions regarding the appropriateness and usage of a single preference-based measure for cancer. The approach taken by cancer measure developers such as the EORTC group [9] and the Functional Assessment of Chronic Illness Therapy group [29] is to provide cancer-specific modules alongside the core measures, which suggests that a single measure is not sufficient, but a module has not been provided in MF.

The valuation sample was younger and healthier and had a higher level of education than did the general population, indicating that it was not representative. This group may have preferences different from those in the general population. The preferred model, however, fits the data well, with low error in predicting health state utility values at the health state level.

The valuation survey design used an orthogonal array to select health states that assumes independence between all dimensions. Although responses to the MF-SAF dimensions were correlated, the assumption of independence was deemed appropriate due to the apparent difference in the underlying symptoms of the dimensions. Respondents in the valuation survey did not indicate that this affected the plausibility of the health states they valued.

A final limitation is that the MF-SAF dimensions have two severity levels for dimensions from the MF-SAF, whereas patients self-report their health using a 0 to 10 scale. This means that preferences have been elicited only for the extreme ends of this

Table 5 – Preference weights for MF-SAF dimensions of the MF-8D by response

Dimension	Response on the 0–10 scale										
	0	1	2	3	4	5	6	7	8	9	10
Method 1: Using equal interval											
Night sweats (or feeling hot and flushed)	0	0.0080	0.0160	0.0240	0.0320	0.040	0.0480	0.0560	0.0640	0.0720	0.0800
Itchiness	0	0.0093	0.0186	0.0279	0.0372	0.0465	0.0558	0.0651	0.0744	0.0837	0.0930
Abdominal discomfort (feel uncomfortable, pressure, or bloating)	0	0.0145	0.0290	0.0435	0.0580	0.0725	0.0870	0.1015	0.1160	0.1305	0.1450
Pain under ribs on the left side	0	0.0139	0.0278	0.0417	0.0556	0.0695	0.0834	0.0973	0.1112	0.1251	0.1390
Bone or muscle pain	0	0.0178	0.0356	0.0534	0.0712	0.0890	0.1068	0.1246	0.1424	0.1602	0.1780
Method 2: Using Rasch item logit thresholds											
Night sweats (or feeling hot and flushed)	0	0.0250	0.0338	0.0410	0.0471	0.0524	0.0573	0.0621	0.0672	0.0731	0.080
Itchiness	0	0.0384	0.0415	0.0441	0.0467	0.0499	0.0541	0.0600	0.0681	0.0789	0.093
Abdominal discomfort (feel uncomfortable, pressure, or bloating)	0	0.0446	0.0608	0.0726	0.0815	0.0888	0.0957	0.1034	0.1134	0.1268	0.145
Pain under ribs on the left side	0	0.0582	0.0645	0.0688	0.0721	0.0758	0.0809	0.0886	0.1001	0.1165	0.139
Bone or muscle pain	0	0.0659	0.0741	0.0819	0.0901	0.0990	0.1095	0.1220	0.1372	0.1556	0.178

MF-SAF, Myelofibrosis-Symptom Assessment Form; MF-8D, Myelofibrosis 8 dimensions.

scale and have been divided equally as well as on the basis of the Rasch logit scale to produce a utility decrement for every point on the 0 to 10 scale, and this may not reflect the preferences of the sample in the valuation survey.

Despite these limitations, results from this study provide a condition-specific preference-based measure that can be used to generate utility values in a population with MF in which the MF-SAF 2.0 (or other versions with similar questions) and the QLQ-C30 have been used. Future research assessing the validity of the classification system of the MF-8D in an external data set including comparisons with generic preference-based measures such as the EQ-5D is recommended.

Acknowledgments

We thank all the clinicians and interviewees who took part in the study as well as the ICON plc researchers who undertook the interviews.

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

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.2015.07.004 or, if a hard copy of article, at www.valueinhealthjournal.com/issues (select volume, issue, and article).

REFERENCES

[1] National Institute for Health and Care Excellence. NICE Guide to the Methods of Technology Appraisal. NICE, London, 2013.
[2] Dolan P. Modelling valuations for EuroQol Health States. Med Care 1997;35:1095–108.
[3] Brazier J, Connell J, Papaioannou D, et al. A systematic review, psychometric analysis and qualitative assessment of generic preference-based measures of health in mental health populations and the estimation of mapping functions from widely used specific measures. Health Technol Assess 2014;18:1–188.
[4] Longworth L, Yang Y, Young T, et al. Use of generic and condition-specific measures of health-related quality of life in NICE decision-making: systematic review, statistical modelling and survey. Health Technol Assess 2014;18:1–224.
[5] Mesa R, Schwager S, Radia D, et al. The Myelofibrosis Symptom Assessment Form (MFSAF): an evidence-based brief inventory to measure quality of life and symptomatic response to treatment in myelofibrosis. Leukemia Res 2009;33:1199–203.
[6] Garau M, Koonal SK, Mason AR, et al. Using QALYs in cancer. Pharmacoeconomics 2011;29:673–85.
[7] Mesa RA, Kantarjian H, Tefferi A, et al. Evaluating the serial use of the Myelofibrosis Symptom Assessment Form for measuring symptomatic improvement: performance in 87 myelofibrosis patients on a JAK1 and JAK2 inhibitor (INCB018424) clinical trial. Cancer 2011;117:4869–77.
[8] 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. Journal of the National Cancer Institute 1993;85:365–76.
[9] Fayers P, Bottomley AO, EORTC Quality of Life Group. Quality of life research within the EORTC—the EORTC QLQ-C30. Eur J Cancer 2002;38:125–33.
[10] Kiladjian J-J, Gisslinger H, Passamonti F, et al. Health-related quality of life (HRQoL) and symptom burden in patients (Pts) with myelofibrosis (MF) in the COMFORT-II study [abstract 6626]. J Clin Oncol 2012;30 (Suppl. 15):6626.
[11] Harrison CN, Messa RA, Kiladjian JJ, et al. Health-related quality of life and symptoms in patients with myelofibrosis treated with ruxolitinib versus best available therapy. Br J Haematol 2013;162:229–39.
[12] Mukuria C, Brazier J, Rafia R. Does the generic cancer outcome measure EORTC QLQ-C30 work in myelofibrosis? Poster presented at: PCN123 ISPOR 20th Annual International Meeting. May 19, 2015, Philadelphia, PA. Value Health 2015;18:A210–1.
[13] Rowen D, Brazier JE, Young TA, et al. Deriving a preference-based measure for cancer using the EORTC QLQ-C30. Value Health 2011;14:721–31.
[14] Roskell NS, Mendelson ET, Whalley D, Knight C. PCN95 Using a condition-specific measure of patient-reported outcomes to derive utilities in myelofibrosis. Value Health 2012;15:A224–5.
[15] Brazier JE, Rowen D, Mavranezouli I, et al. Developing and testing methods for deriving preference-based measures of health from condition-specific measures (and other patient-based measures of outcome). Health Technol Assess 2012;16:1–114.

[16] Verstovsek S, Mesa RA, Gotlib J, et al. A double-blind, placebo-controlled trial of ruxolitinib for myelofibrosis. N Engl J Med 2012;366:799–807.
[17] Lloyd A, Kerr C, Breheny K, et al. Economic evaluation in short bowel syndrome (SBS): an algorithm to estimate utility scores for a patient-reported SBS-specific quality of life scale (SBS-QoL™). Qual Life Res 2014;23:449–58.
[18] Mulhern B, Rowen D, Brazier J, et al. Development of DEMQOL-U and DEMQOL-PROXY-U: generation of preference-based indices from DEMQOL and DEMQOL-PROXY for use in economic evaluation. Health Technol Assess 2013;17: v–140.
[19] Mulhern B, Rowen D, Jacoby A, et al. The development of a QALY measure for epilepsy: NEWQOL-6D. Epilepsy Behav 2012;24:36–43.
[20] Field A. Discovering Statistics Using SPSS (2nd ed.). London, England: Sage, 2005.
[21] Young T, Yang Y, Brazier JE, et al. The first stage of developing preference-based measures: constructing a health-state classification using Rasch analysis. Qual Life Res 2009;18:253–65.
[22] Young T, Yang Y, Brazier J, Tsuchiya A. The use of Rasch analysis in reducing a large condition-specific instrument for preference valuation: the case of moving from AQLQ to AQL-5D. Med Decis Making 2011;31:195–210.
[23] Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. Chicago: University of Chicago Press, 1960.
[24] Tesio L. Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. J Rehabil Med 2003;35:105–15.
[25] IBM Corp. IBM SPSS Statistics for Windows, Version 21.0. Armonk, NY: IBM Corp., 2012.
[26] Devlin N, Tsuchiya A, Buckingham KJ, Tilling C. A uniform time trade off method for states better and worse than dead: feasibility study of the 'lead time' approach. Health Econ 2011;20:348–61.
[27] Kind P, Hardman G and Macran S. UK Population Norms for EQ-5D. Centre for Health Economics. 1999 Discussion Paper No 172.
[28] Oken MM, Creech RH, Tormey DC, et al. Toxicity and response criteria of the Eastern Cooperative Oncology Group. Amer J Clin Oncol 1982;5:649–56.
[29] Webster K, Cella D, Yost K. The Functional Assessment of Chronic Illness Therapy (FACIT) measurement system: properties, applications, and interpretation. Health Qual Life Outcomes 2003;1:79.