
Monograph:
Rowen, D., Brazier, J.E., Young, T.A. et al. (4 more authors) (2010) Deriving a preference-based measure for cancer using the EORTC QLQ-C30. Discussion Paper. (Unpublished)

HEDS Discussion Paper 10/01

HEDS Discussion Paper 10/01


ScHARR
School of Health and
Related Research

Deriving a preference-based measure for cancer using the EORTC QLQ-C30

Rowen D[1*], Brazier JE[1], Young TA[1], Gaugrist S[2], Craig BM[3,4], King MT[5], Velikova G[6]

1. School of Health and Related Research, University of Sheffield, Sheffield S1 4DA, UK
2. Janssen-Cilag Ltd, High Wycombe, UK
3. Health Outcomes & Behavior, Moffitt Cancer Center, USA
4. Department of Economics, University of South Florida, USA
5. Psycho-oncology Co-operative Research Group (PoCoG), School of Psychology, University of Sydney, Australia
6. Cancer Research UK Clinical Centre, Leeds, UK

* Correspondence to: Donna Rowen, Health Economics and Decision Science, University of Sheffield, Regent Court, 30 Regent Street, Sheffield S1 4DA, UK
Telephone: +44 (0)114 222 0728
Fax: +44 (0)114 272 4095
E-mail: d.rowen@sheffield.ac.uk

Abstract

Background: The EORTC QLQ-C30 is one of the most commonly used measures in cancer but in its current form cannot be used in economic evaluation as it does not incorporate preferences.

Methods and results: We address this gap by estimating a preference-based single index for cancer from the EORTC QLQ-C30 for use in economic evaluation. Factor analysis, Rasch analysis and other psychometric analyses were undertaken on a clinical trial dataset of 655 patients with multiple myeloma to derive a health state classification from the QLQ-C30 that is amenable to valuation. The resulting health state classification system has 8 dimensions (physical functioning, role functioning, social functioning, emotional functioning, pain, fatigue and sleep disturbance, nausea, and constipation and diarrhoea) with 4 or 5 levels each. A valuation study was conducted of 350 members of the UK general population using ranking and time trade-off. Mean and individual level additive multivariate regression models including the episodic random utility model were fitted to the valuation data to derive preference weights for the classification system. Mean absolute error ranges from 0.046 to 0.054 and models have few inconsistencies (0 to 2) in estimated preference weights.

Conclusions: We conclude that it is feasible to derive a preference-based measure from the EORTC QLQ-C30 for use in economic evaluation, but this work needs to be extended to other countries and replicated across other patient groups.

Key words: Preference-based measures; QALYs; EORTC QLQ-C30

* * * * *

Acknowledgements: We would like to thank the Centre for Research and Evaluation at Sheffield Hallam University for conducting the interviews. The studies reported in this paper were funded by Janssen-Cilag Ltd.

Introduction

Generic preference-based measures, such as the EQ-5D [1], HUI3 [2] or SF-6D [3] are widely used to calculate quality-adjusted life years (QALYs) [4] for use in economic evaluation. The EQ-5D is the most common generic preference-based measure (PBM) and is currently recommended by NICE [5]. However, generic measures of health have been found to be inappropriate or insensitive for some medical conditions [6] and for cancer in particular [7]. Furthermore, clinicians and researchers often choose to include condition-specific measures such as the EORTC QLQ-C30 in trials rather than generic preference-based measures because they need measures which are sensitive to the effects of interventions across a range of relevant symptoms, side effects and aspects of functioning and quality of life. Condition-specific measures, such as the EORTC's core quality of life questionnaire, the QLQ-C30, have great clinical utility because they summarise a number of symptom and domain-specific scales. However, because they are not preference-based, they provide a description rather than a valuation of health, and therefore cannot be used to estimate QALYs.

There are three ways in which researchers can estimate utilities to produce QALYs for a trial where a generic preference-based measure such as the EQ-5D is unavailable: undertake 'mapping'; value vignettes that describe the health states covered by patients in the trial; or derive a preference-based measure from the existing condition-specific measure.

Mapping (also known as 'cross-walking' or estimating exchange rates between instruments) involves predicting the relationship between the non-preference-based measure, for example the EORTC QLQ-30, and a generic preference-based measure, for example the EQ-5D, using statistical association. Mapping by statistical association may be considered less arbitrary than using the judgement of experts to map between measures. Typically mapping by statistical association uses two datasets; an estimation dataset that contains respondents' self-reported scores for their own health using, for example, EORTC QLQ-C30 and EQ-5D and the study dataset that contains only EORTC QLQ-C30. A statistical relationship between the EORTC QLQ-C30 and EQ-5D is estimated using regression techniques on the estimation dataset and the results are applied to the study dataset to obtain predicted EQ-5D health state utility values. Mapping is a second best alternative to the use of a preference-based measure directly in the study as mapped estimates can have large errors, most noticeably when mapping from condition-specific measures to generic preference-based measures [8]. Mapping requires: 1) a degree of overlap between the descriptive systems of both measures, 2) the relationship estimated in the estimation dataset is generalisable to the study dataset, and 3) both measures are administered on the same population. Yet this means that mapping is valid only if both measures are appropriate for the patient population, which is unlikely to be the case for generic preference-based measures administered to cancer patients [7].

An alternative solution to estimate utilities when generic preference-based measures are unavailable or inappropriate is to value a selection of bespoke descriptions or vignettes that describe the health states covered by patients in the trial. However, patients in the trial experience a variety of different health states, and as only a selection will be valued this will not fully take into account variation across individuals and across treatments. It is also unlikely that identical health states are experienced in trials for different treatments and hence vignettes need to be created and valued for each trial which reduces comparability across trials and treatments.

It has therefore been argued that a better approach in many cases would be to develop a preference-based measure from the condition-specific questionnaire specifically designed for that condition [9]. Typically this requires reducing the length of the questionnaire to obtain a health state classification system that remains responsive and valid for the condition but that is amenable to valuation. Preference weights for the health state classification are then obtained from a representative sample of the general population.

This study applies the methods originally developed in the estimation of a generic preference-based measure of health from the SF-36 [3, 10] and subsequently used with condition specific measures in urinary incontinence [11], asthma [12, 13] and overactive bladder [14, 15]. The study involved three stages. First, a health state classification system was derived from the EORTC QLQ-C30 that is amenable to valuation using a recognised preference elicitation technique. The classification system was derived using psychometric analysis, factor analysis and Rasch analysis on a dataset of patients with multiple myeloma. Second, a valuation survey was conducted asking members of the general population to value a sample of states defined by the classification. Third, regression models were estimated on the results of the valuation survey to estimate the preference weights to produce a utility estimate for every health state defined by the classification system.

EORTC QLQ-C30

The EORTC QLQ-C30 is one of the most commonly used measures in cancer [16] and dominates cancer clinical trials in Europe and Canada. The EORTC QLQ-C30 contains 30 questions that cover the most common symptoms of cancer (such as pain, fatigue, nausea and vomiting) and various aspects of function (including physical, role, social, emotional and cognitive functioning). The EORTC QLQ-C30 is summarised using fourteen scales, each representing a particular symptom or aspect of function, plus one global quality of life scale (based on two global questions). Its validity has been well established for many conditions in cancer.

While the EORTC QLQ-C30 has proved to be a useful instrument for demonstrating treatment benefits, it cannot be used in economic evaluation in

its current form because it does not incorporate preference information. While it generates a profile of scores representing a range of symptoms and aspects of functioning and a global quality of life score, it does not currently generate a single preference-based index of quality of life required for economic evaluation using QALYs. Further, the number of items and scales is too large to be amenable to valuation using preference elicitation techniques such as time trade off and standard gamble.

EORTC QLQ-C30 multiple myeloma patient dataset

The dataset used to derive the health state classification system contains data on patients newly diagnosed with multiple myeloma. Data were collected in VISTA (Velcade as Initial Standard Therapy in Multiple Myeloma: Assessment with Melphalan and Prednisone), a phase III randomized open-label trial (ClinicalTrials.gov number, NCT00111319) completed in June 2007. Patients were asked to complete the EORTC QLQ-C30 at screening visit, day 1 of each cycle of treatment (cycles 1-9), end of treatment visit and in post treatment phase (every 6 or 8 weeks) until disease progression. The screening phase of the dataset (n=655) is used to select items for the health state classification. Data at cycle 5 of treatment (n=471) in the trial are then used to validate the choice of items for a different time period where responses are likely to have changed.

Methods

Methodology used to derive a health state classification from the QLQ-C30

The aim was to produce a multidimensional health state classification from the EORTC QLQ-C30 that is amenable to valuation by respondents with a minimum loss of information and subject to the constraint that responses to the original instrument can be unambiguously mapped onto it. This implies that the text of the items should be altered as little as possible. The task is therefore to determine the dimensions, items and the levels of the EORTC QLQ-C30 to be included in the new classification. The methodology outlined here uses a combination of Rasch and classical psychometric analysis [15]. SPSS version 15 [17] was used for the factor analysis and Rasch Unidimensional Measurement Models (RUMM2020) [18] was used for the Rasch analysis.

Dimensional structure

Multidimensional health state classifications for valuation should have structural independence between dimensions to avoid nonsensical states [2]. In other words, there must be little correlation between the dimensions in our classification system. The large literature on the EORTC QLQ-C30 focuses upon its use as a profile measure of health, but here we wish to determine the dimensions across all items, ignoring whether these are functions or symptoms.

Factor analysis can be used for a set of observed variables to identify structurally independent dimensions by highlighting underlying factors that

explain patterns of correlation [19]. We applied factor analysis to 27 of the 30 items of the EORTC QLQ-C30, (excluding global quality of life and financial impact items as these are inappropriate for a PBM of health-related quality of life) to explore the dimension structure of the EORTC QLQ-C30. The dimensions were determined using a varimax component matrix and eigenvalues. The extent to which items belong to a single dimension can also be examined using Rasch analysis (see below). The results were discussed with our team's clinical expert (GV) to make sure that they made sense clinically before making a final decision on the dimensionality of the health state classification.

Item selection

Each dimension of a health state classification system of a preference-based measure is usually represented by just one or two items to render the system amenable to preference elicitation methods. We used the following conventional psychometric criteria to help select items from the QLQ-C30: distribution of responses across categories of response (including floor effects and ceiling effects), the percentage of missing data, correlation of item to dimension, and responsiveness to change over two points in time using the standardised response mean (SRM).

A further technique often used [13, 15, 20, 21] to select items is Rasch analysis [22]. This is a mathematical technique that converts qualitative (categorical) responses to points on a continuous (unmeasured) latent scale using a logit model [23]. It can be used to assess whether an item fits the model, the severity of health problem being covered by each item, the extent to which items have response choices that are appropriately ordered as responders can distinguish between the response levels for a given item and whether items perform differently between populations (known as differential item functioning (DIF)) [15, 20]. Items that fit the Rasch model, cover the full range of severity, have ordered response choices and do not suffer from DIF were considered as candidates for inclusion in the health state classification. Items that did not fit the Rasch model (using criteria that item level Chi-square P-value<0.01) were removed; all other items were retained and the Rasch analysis was re-estimated. We used the following criteria to assess the Rasch model for each dimension: item-trait interaction (whether data fitted the Rasch model for groups of respondents with similar underlying health); person separation index (PSI) (whether the Rasch model could discriminate between responders); item fit and person fit residuals (the divergence between expected and observed responses per respondent); and item range and spread at logit zero (whether items covered a wide range of severity).

The final selection of dimensions, items and levels for the health state classification was based on what appeared to perform best using the psychometric tests and Rasch analysis and at the same time ensuring that health states made clinical sense and were amenable to valuation by respondents. The process involved judgment by our clinical expert (GV) and

consideration of other factors such as wording of the health state classification system.

Valuation study to obtain preferences for the health state classification

The second stage of the research was to obtain valuations of states defined by the health state classification system. Key methodological issues were the choice of technique for eliciting preferences, the sample of states valued, sampling of respondents and overall size of sample.

The valuation technique used to value health states was Time Trade-off (TTO). States were sampled using an orthogonal array in order to enable the estimation of an additive model for the preference weights. Use of an orthogonal array to sample states is common when dimensions are independent. SPSS version 15 was used to produce a sample of 81 states using orthogonal array and this was supplemented by 4 additional states. We chose to value 85 states in order to enable each respondent to value the worst state, an equal number of responses per state, and an equal number of states to be valued per respondent.

At the interview, respondents read the descriptive system and self-completed the system for their own health. The EORTC QLQ-C30 does not mention cancer in its descriptive system and therefore respondents were not aware that the health states were used to describe the health-related quality of life of cancer patients. Respondents were asked to rank 8 states alongside 'full health' and 'dead' to help familiarize them with the states. Respondents then valued the same 8 states using the Measurement and Valuation of Health (MVH) study version of TTO that involves the use of a visual prop designed by the MVH group (University of York) [24]. Respondents were initially taken through a hypothetical TTO to help them understand the task. For each health state respondents were first asked whether they would prefer the given health state for 10 years after which they will die, or to die immediately. For health states considered better than being dead (BTD), respondents were asked to choose between (a) health state *h for w* years, after which they will die, or (b) full health for *z* years ($z \leq w$), after which they will die. While *w* is fixed at 10 years, years in full health, *z*, is varied to determine the point where respondents are indifferent between the two options. For health states considered worse than being dead (WTD) respondents were asked to choose between (a) health state *h* for *w* years followed by full health for *z* years after which they will die, or (b) immediate death. Both years in optimal health, *z*, and years in health state, *w*=10-*z*, are varied to determine the point where respondents are indifferent between the two options. After the collection of trade-off responses (*w, z*), respondents were asked a number of background questions covering demographic and socio-economic characteristics.

A representative sample of the general population was interviewed in their own homes by trained and experienced interviewers who had worked on numerous previous valuation surveys, such as the HUI2 [25] and OAB-5D [14].

The valuation data

Respondents were from geographical areas in the North of England including urban and rural areas with a mix of socio-economic characteristics. There were 350 successfully conducted interviews, a response rate of 40.3% for suitable respondents answering their door at time of interview. Six respondents were excluded from the analysis; three were excluded for valuing all states as identical and less than one; two respondents were excluded for valuing the worst possible health state higher than every other state; one respondent was excluded for valuing all states as worse than dead. All other responses were used in the analysis reported here. The remaining 344 respondents had a health state completion rate of 98.5% in the TTO tasks. Characteristics of the included respondents are compared to the general population in South Yorkshire and England (Table 1). The valuation sample contained a higher proportion of people aged 41-65, retired people, females and people in poorer health than in the population at large.

Modelling to obtain preference weights for the health state classification

The TTO responses (*z*,*w*) were analysed using a range of different specifications. The standard specification is based on the approach first used for the UK EQ-5D preference weights [24]:

$$y_{ij}^{s} = f(\mathbf{X}_{\delta\lambda}\boldsymbol{\beta}) + \varepsilon_{ij}^{s}, \quad y_{ij}^{s} = \begin{cases} 1 - z_{ij}/w_{ij} & \text{if state better than dead} \\ 1 + z_{ij}/10 & \text{if state worse than dead} \end{cases} \qquad (1)$$

where *i*=1,2 ...*n* represents individual health states and *j*=1,2...*m* represents respondents. The dependent variable y_{ij}^{s} is disvalue 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 health state classification. Level λ = 1 acts as a baseline for each dimension.

The second specification is the episodic random utility model (ERUM), where the value of the health state depends on its duration, w_{ij}:

$$y_{ij}^{e} = w_{ij} f(\mathbf{X}_{\delta\lambda}\boldsymbol{\beta}) + \varepsilon_{ij}^{e}, \; y_{ij}^{e} = \begin{cases} w_{ij} - z_{ij} & \text{if state better than dead} \\ w_{ij} + z_{ij} & \text{if state worse than dead} \end{cases} \qquad (2)$$

In order to produce error terms on the same scale as the standard specification in equation (1), both z_{ij} and w_{ij} are divided by 10 before estimation.

The standard approach transforms WTD TTO responses to bound value estimates at -1. This has been criticised, as there is little empirical evidence for why values should be bound at -1 and arguably the transformed responses cannot be interpreted as being measured on the same utility scale as states BTD [26]. The ERUM model was developed to deal with this criticism by changing the way TTO responses are modelled. Under the ERUM, WTD TTO

responses are not transformed and are therefore modelled in a way that is consistent with BTD responses [27].

Each model was estimated using ordinary least squares (OLS) estimation with clusters at the respondent-level, which assumes that responses may be correlated within respondent and are independent between respondents. Random and fixed effects were estimated in the standard approach to take into account individual differences in values (Brazier et al, 2002). For the random and fixed effects model the error term, ε_{ij} is subdivided as follows:

$$\varepsilon_{ij} = u_j + e_{ij} \qquad (3)$$

where u_j is individual random effect and e_{ij} is the random error term for the *i*th health state valuation of the *j*th individual. Maximum likelihood estimation was used for the random effects estimation.

Instead of examining individual responses, the dependent variable under the standard specification, y_{ij}^s, may be condensed into mean estimates, $\bar{y}_i$, for each *i*th state. Using the 85 mean estimates, the standard model (equation 1) can also be estimated via OLS:

$$\bar{y}_i = f(\mathbf{X}_{\delta k}\boldsymbol{\beta}) + \varepsilon_i \qquad (4)$$

The use of mean estimates diminishes the effects of outliers in the distribution of y_{ij}^s. The impact of adding interaction terms and socio-demographic variables is explored for all models.

Several alternative criteria to indicate model performance are reported. The difference between actual and predicted values is assessed using mean absolute error (MAE) calculated at the health state level. MAE is an indicator of how large the prediction errors are and reporting the number of health states valued with errors greater than 5% and 10% indicates whether the errors are of a minimal important difference. Inconsistencies in parameter estimates for adjacent levels of an item were noted as these indicated that worsening health did not lead to a lower utility value. Models in which these inconsistencies were removed by merging levels were also estimated. The number of main effects with insignificant coefficients is also reported. Performance of all regression models is reported using inconsistencies, significant coefficients, mean absolute error of health state predictions and MAE greater than 5% and 10% and by examining plots of actual and predicted health state values.

Results

Health state classification

Step 1: Instrument dimensionality

A four factor model on all 27 items accounted for 58.7% of the variance. Items were divided into the four factors according to their 'loadings', the correlation between the item and the factor. All items loaded >0.35 on a factor, but some

items cross loaded. Factor 1 contained the majority of items (14), covering physical functioning, role functioning, social functioning, pain and items 'need a rest', 'felt weak' and 'difficulty concentrating'. Factor 2 contained all items covering emotional functioning plus 'trouble sleeping'. Factor 3 contained all items covering eating and digestion and factor 4 contained items 'short of breath', 'were you tired' and 'difficulty remembering'. Items loading into factors 2 and 3 were conceptually clear and in accordance with the grouping of the original EORTC instrument. Items loading into factors 1 and 4 (17 items), on the other hand, covered a large range of concepts and further factor analysis was done on these items to determine whether further differentiation was possible. The additional item 'trouble sleeping' was added (18 items in total) over a concern that the initial factor analysis captured some causality rather than correlation.

A four factor model on the 18 items explained 67.6% of the variance. All items loaded >0.4 on a factor, but some items cross loaded (difference between cross loadings <0.2). Cross loading items were included in the factor that was clinically meaningful. The four factors were: physical functioning, role functioning and pain; social functioning; fatigue, trouble sleeping and short of breath; cognitive functioning (principal-component rotated factor loadings are available from authors on request).

Table 2 shows the potential items categorised according to the dimensions for consideration for the measure amenable to valuation. Overall the 27 items can be divided into 6 factors or dimensions: (1) physical functioning, role functioning and pain, (2) social functioning, (3) emotional functioning, (4) digestion, (5) fatigue, trouble sleeping and shortness of breath and (6) cognitive functioning. After consultation with our clinical specialist (GV), cognitive functioning (items 20 and 25) and shortness of breath (item 8) were excluded from the PBM on the grounds that they are neither a symptom nor side effect of treatment for multiple myeloma, leaving 24 items remaining. The remaining five dimensions are used for the following analysis to determine the PBM, where items were fitted to Rasch models for each of the five dimensions.

Step 2: Selecting items by dimension

Table 2 presents psychometric analysis and goodness of fit for the Rasch models for each dimension.

Item-level ordering and differential item functioning

Only item 15 (digestion dimension) was disordered and for this item 'a little', 'quite a bit' and 'very much' were collapsed into one level prior to proceeding with further Rasch modelling. Four items demonstrated differential item functioning by sex and were split according to sex: items 1 and 3 (physical functioning, role functioning and pain dimension), item 15 (digestion) and item 22 (emotional functioning). Item 21 (emotional functioning) demonstrated differential item functioning by age and was split according to age. This indicated that these items were not ideal for the health state classification.

Rasch model goodness of fit
Three items demonstrated poor item fit (Chi-square P-value<0.01) and were excluded from subsequent Rasch models estimated for each dimension: item 6 (physical functioning, role functioning and pain dimension), item 14 (digestion), item 11 (fatigue).

Physical functioning, role functioning and pain
This dimension covers three separate attributes of quality of life: physical functioning, role functioning and pain. Each item covers only one attribute and it is unlikely that an item on physical functioning, for example, will capture or reflect role functioning or pain. In order to accurately represent the entire dimension, and in accordance with the EORTC QLQ-C30 scaling and with clinical opinion, we decided that a minimum of one item each for physical functioning, role functioning and pain was required.

Out of the 5 items that capture physical functioning, items 4 and 5 did not capture the full range of severity but had the highest SRM suggesting better responsiveness and ability to capture severe health problems, though in all cases they were low according to Cohen's criteria [28]. Item 2 overall performed well with relatively high item fit but had floor effects, relatively low SRM and limited range of severity. Figure 1 shows the item map for all items in this dimension, reporting three thresholds between response categories 'not at all' and 'a little', 'a little' and 'quite a bit', and 'quite a bit' and 'very much'. No item captured the full severity range in terms of coverage as items 3, 4 and 5 captured severe health whereas items 1 and 2 captured less severe health. Given that none on the items were ideal, items 2 and 3 were chosen in order to cover the full severity range and there is a coherence since both measure trouble walking. Items 2 and 3 were merged to a five level item in the health state classification, with levels 1 to 4 taken directly from item 2 (no/a little/quite a bit/very much trouble taking a long walk) and level 5 (very much trouble taking a short walk outside of the house) taken from level 4 of item 3.

Of the two role functioning items, item 7 had ceiling effects, but also had good Rasch model item fit and relatively good item range (Figure 1). Item 6 performed similarly to item 7 for the psychometric analysis, but had poor Rasch model item fit. Therefore item 7 was selected for the health state classification to represent role functioning.

Items 9 and 19 capture pain. Item 9 had a large severity range, but poor item fit, very high item fit residual and low item level p-value suggesting the item contributes poorly to the dimension. Item 19 had a more limited range and more evidence of ceiling effects but better item fit. Item 19 was chosen due to better item fit and due to its wording, since it measures the extent to which pain interferes with daily activities rather than the existence of pain *per se*, which is likely to be more important to respondents.

Social functioning
Of the two social functioning items, item 26 had a slightly higher degree of ceiling effects, higher item fit residual and lower p-value. Further, it may not be applicable to all respondents as it captures whether physical condition or medical treatment interferes with family life. Item 27 was chosen as it performs marginally better psychometrically and is arguably applicable to a higher number of respondents as it measures interference with social activities.

Emotional functioning
Items 21 and 22 suffered from DIF. Items 23 and 24 performed similarly across all criteria. Item 23 had a larger severity range than item 24, but also had a higher item fit residual, suggesting greater divergence between expected and observed responses. Item 24 was chosen as it overall performs best, had a higher SRM and was felt to be more clinically relevant to patients with cancer.

Digestion
None of the digestion items performed well in psychometric or Rasch analysis. Despite this, they were included in the health state classification as it was felt that they were clinically important for patients with multiple myeloma. These items capture multiple symptoms of digestion-related problems: lack of appetite, nausea, vomiting, constipation and diarrhoea. Lack of appetite, nausea and vomiting (items 13, 14, 15) are all closely related symptoms, and constipation and diarrhoea (items 16 and 17) are bowel symptoms, suggesting the items can be separated into two attributes. Item 13 was the only item from appetite, nausea and vomiting that did not suffer from DIF, item level disordering or poor item fit. However, this item performed poorly in the Rasch model with small coverage at logit zero and high item fit residuals. After consultation amongst our research team, including our clinical specialist, it was decided that item 13, which captures lack of appetite, would not be chosen because it may be thought to be a desirable (positive) symptom by some people and may be a symptom due to the age of the population rather than the condition. Therefore, despite suffering from problems in the Rasch model items 14 and 15 were considered as it was felt to be important to capture appetite, nausea or vomiting in the health state classification. Both items suffer from large ceiling effects and have low SRM, with item 14 performing marginally better. Therefore item 14 (nausea) was chosen for the health state classification.

Items 16 and 17 on constipation and diarrhoea perform similarly, both suffering from extreme ceiling effects, small spread at logit zero and high residuals. It was decided to combine these items as they are both bowel symptoms where respondents rarely suffer from both during a weekly period. The items were combined such that level 1 of the merged item captures no bowel (constipation or diarrhoea) problems, levels 2, 3 and 4 capture ‘a little’ ‘quite a bit’ and ‘very much’ constipation and/or diarrhoea.

Fatigue and sleep disturbance
Items 10, 12 and 18 performed similarly in the Rasch and psychometric analyses, with large severity range and no large ceiling or floor effects. However, item 10 had a relatively high item fit residual and item 12 has a low item p-value, indicating it does not contribute well to the dimension in Rasch models. Item 18 performed marginally better, with a relatively high p-value, low residual, large coverage at logit zero and large range, and so was selected for the health state classification.

Health state classification
The classification system has 8 dimensions (physical functioning, role functioning, pain, emotional functioning, social functioning, fatigue and sleep disturbance, nausea, constipation and diarrhoea) made up of 10 items. Table 4 summarises the items chosen from the EORTC QLQ-C30 for each dimension of the health state classification. Table 5 presents the final health state classification system of dimensions and their levels. A health state is made up of 8 sentences and hence has an 8 digit identifier, from best state 11111111 to worst state 54444444. This system generates a total of 81,920 health states.

Valuation survey
Mean TTO values varied from 0.95 for best state to 0.13 for worst state (available from authors on request), which suggests that on average all states were valued as better than being dead. Of the total 2710 TTO observations, 514 observations (19%) were equal to 1 (equivalent in value to full health) and 271 (10%) were less than or equal to 0 (valued as the same or worse than being dead).

Modelling health state values
Table 6 presents the preference weights estimated using a variety of regression models. All coefficients have the expected sign (i.e., level 1 on each dimension is the reference point and higher levels increase TTO disvalue), their size is consistent with the severity scale in all but two cases (i.e., higher levels have larger coefficients and an increasing increment on TTO disvalue except physical functioning levels 4 and 5 and nausea levels 2 and 3) and the majority of coefficients are statistically significant. Models (1), (2), (4) and (5) are based on the standard specification outlined in equations (1) and (4), model (3) uses the ERUM specification in equation (2). Models (1) and (3) are individual level models estimated using OLS, model (2) is a random effects model estimated using maximum likelihood. The results of the Hausman test confirmed that a fixed effects model would render similar estimates at reduced efficiency. Models (4) and (5) are mean level models. Model (5) is a consistent version of model (4) where adjacent inconsistent levels are merged into a common dummy variable.
Mean absolute error was similar between models, ranging from 0.046 to 0.054. The number of health states with errors greater than 5% ranges from 33 to 41 and errors greater than 10% ranges from 6 to 13. Models including interaction effects and socio-demographic were estimated (available from the

authors by request) but predictive ability, inconsistencies and significant coefficients for the main effects variables were not improved.

Discussion

We have estimated a preference-based measure for the EORTC QLQ-C30 using methodology first applied in the development of the SF-6D from the SF-36 [3] and a number of condition specific measures. The derived preference-based measure including 8 dimensions, the EORTC-8D, was constructed using psychometric analysis (including Rasch) to ensure that chosen items appropriately reflected their dimension and that each dimension covered a wide range of severity. A sample of states was valued using time trade-off and then modelled using a variety of specifications. The estimated preference weights enable utility scores to be generated directly from EORTC QLQ-C30 datasets.

An important concern is that often condition-specific measures fail to capture co-morbidities and side-effects of treatments, and hence are not strictly comparable to generic measures when used to estimate QALYs for resource allocation. One way to enhance comparability across measures is for all measures to use the same methodology to derive values [6]. Our valuation study followed the methodology used in the development of generic measures: we implemented the protocol used to derive the UK EQ-5D preference weights [24]; used common anchors of 1 for full health and zero for dead; and interviews were conducted using a sample of the general population. Furthermore, the EORTC-8D descriptive system captures a wide range of dimensions including generic dimensions such as physical functioning and role functioning, as well as more condition-specific symptoms such as nausea, constipation and diarrhoea. Therefore the descriptive system is likely to capture overall health-related quality of life including both comorbidities and side-effects. An area of future research will be to compare this cancer-specific measure to generic measures in terms of sensitivity and validity.

A general concern regarding the development of PBMs from existing questionnaires is that the classification system is strongly influenced by the specific patient dataset used to develop the classification. While the patient dataset is international, the valuation study used here is a UK dataset and hence the EORTC-8D classification and preference weights presented here might be more appropriate for UK trials. The health state classification was developed using data in newly diagnosed multiple myeloma patients. Further testing of the classification will be undertaken across datasets of cancer patients with different types of cancers. This study forms part of a wider cross-country study that will examine the use of preference-based measures from the EORTC QLQ-C30 on a variety of countries and different patient groups.

Given that the health state classification measures cancer, it is somewhat surprising that all states have a positive mean TTO value and hence at the aggregate level all states are valued as being better than being dead. This is in contrast to other valuation studies such as the UK EQ-5D valuation study where 38% (16/42) of health states valued were on average valued as being worse than dead (with mean TTO below zero). One hypothesis is that respondents would view the states differently if cancer was included in the health state description, and this is a topic for future research.

The preference weights should have no inconsistencies; health state values should always decrease as health states become more severe. Models based on the standard specification produced inconsistencies for physical functioning levels 4 and 5. This may have been due to the merging of items in the EORTC QLQ-C30 to form this dimension. Models (1) and (3) also produce inconsistencies for nausea levels 2 and 3. Model (5) removes inconsistencies by merging variables. In comparison, models (3) and (5) perform best overall according to the criteria of predictive ability, inconsistencies and significant coefficients, with model (3) performing better using all criteria. Model (5) has a predicted range of utilities from 1 to 0.199 whereas model (3) has a predicted range of 1 to 0.291, meaning that the worst state defined by the classification has a much lower value using preference weights estimated using model (5). Deciding upon the preferred model comes down to a choice between the mean model (5) with no inconsistencies, as chosen both in the valuation of the SF-6D [29] and the overactive-bladder-specific measure [14], or the recently developed ERUM specification. Although model (5) is in accordance with the recommended value set of many similar measures, the ERUM model (3) is here the preferred model as it more appropriately deals with TTO values for SWD and performs best.

The EORTC-8D was developed out of a concern that generic measures were not appropriate to measure the quality of life of cancer patients [7]. The EORTC-8D enables QALYs to be directly estimated using the EORTC QLQ-C30, a questionnaire typically included in cancer trials, rather than the use of generic measures that are less appropriate [7] or mapping to generic measures that is both less appropriate and increases error around utility estimates. It is hoped that this measure will provide appropriate and useful information for cost per QALY analysis undertaken in cancer trials.

References

1. Brooks R 1996.
EuroQol Group. EuroQol: the current state of play.
Health Policy 3:53-72.

2. Feeny D, Furlong W, Torrance GW, Goldsmith CH, Zhu Z, DePauw S, Denton M, Boyle M (2002).
Multiattribute and single attribute utility functions for the Health Utilities Index Mark 3 system.

Medical Care 40:113-28.

3. Brazier J, Roberts J, Deverill M (2002).
The estimation of a preference-based single index measure for health from the SF-36.
Journal of Health Economics 21(2):271-92.

4. Drummond M (1994).
Economic evaluation alongside clinical trials.
London: Department of Health.

5. NICE (2008).
Guide to the methods of technology appraisal.
http://www.nice.org.uk/aboutnice/howwework/devnicetech/technologyappraisalprocessguides/guidetothemethodsoftechnologyappraisal.jsp
London.

6. Brazier JE, Ratcliffe J, Tsuchiya A, Solomon J (2007).
Measuring and valuing health for economic evaluation.
Oxford: Oxford University Press.

7. Garau M, Shah K, Towse A, Wang Q, Drummond M, Mason A (2009).
Assessment and appraisal of oncology medicines: does NICE's approach include all relevant elements? What can be learnt from international HTA experiences?
Report for the Pharmaceutical Oncology Initiative (POI)

8. Brazier J, Yang Y, Tsuchiya A, Rowen D (2009).
A review of studies mapping (or cross walking) from non-preference based measures of health to generic preference-based measures.
European Journal of Health Economics, in press.

9. Brazier JE, Dixon S (1995).
The use of condition specific outcome measures in economic appraisal.
Health Economics 4:255-64.

10. Brazier JE, Harper R, Thomas K, Jones N, Underwood T (1998).
Deriving a preference based single index measure from the SF-36.
Journal of Clinical Epidemiology 51(11):1115-29.

11. Brazier JE, Czoski-Murray C, Roberts J, Brown M, Symonds T, Kelleher C (2008). Estimation of a preference-based index from a condition specific measure: the King's Health Questionnaire.
Medical Decision Making 28(1):113-26.

12. Yang Y, Tsuchiya A, Brazier J, Young T (2007).
Estimating a preference-based single index from the Asthma Quality of Life Questionnaire (AQLQ).

Health Economics and Decision Sciences Discussion Paper 07/02.
http://www.shef.ac.uk/scharr/sections/heds/discussion.html

13. Young T, Yang Y, Brazier J, Tsuchiya A (2007).
The use of Rasch analysis as a tool in the construction of a preference based measure: the case of AQLQ.
Health Economics and Decision Sciences Discussion Paper 07/01.
http://www.shef.ac.uk/scharr/sections/heds/discussion.html

14. Yang Y, Brazier JE, Tsuchiya A, Coyne K (2009).
Estimating a preference-based index from the Overactive Bladder questionnaire.
Value in Health 12(1):159-66.

15. Young T, Yang Y, Brazier J, Tsuchiya A, Coyne K (2009).
The first stage of developing preference-based measures: constructing a health-state classification using Rasch analysis.
Quality of Life Research 18:253-65.

16. Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, et al (1993).
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 85(5):365-76.

17. SPSS for Windows. Release. 14.0.1. 2005. Chicago: SPSS Inc. 2005

18. Rasch Unidimensional Measurement Models (RUMM) 2020. RUMM Laboratory Pty Ltd 1997-2004.

19. Chatfield C, Collins AJ (1980).
Introduction to Multivariate Analysis.
Chapman and Hall.

20. Young T, Rowen D, Brazier J, Norquist J, Ambegaonkar B, Sazonov V (2009).
Developing preference-based health measures: using Rasch analysis to generate health state values.
Health Economics and Decision Sciences Discussion Paper, forthcoming.

21. Mavranezouli I, Brazier JE, Young TA, Barkham M (2009).
Using Rasch analysis to form plausible health states amenable to valuation: the development of CORE-5D Utility from CORE-OM in order to elicit preferences for common mental health problems.
Health Economics and Decision Sciences Discussion Paper 09/09.
http://www.shef.ac.uk/scharr/sections/heds/discussion.html

22. Rasch G (1960).
Probabilistic models for some intelligence and attainment tests.

Chicago: University of Chicago Press. Reprinted 1980.

23. Tesio L (2003).
Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research.
Journal of Rehabilitation Medicine 35(3):105-15.

24. Dolan P (1997).
Modelling valuations for EuroQol health states.
Medical Care 35(11):1095-108.

25. McCabe C, Stevens K, Roberts J, Brazier J (2005).
Health state values for the HUI 2 descriptive system: Results from a UK survey.
Health Economics 14:231-44.

26. Patrick DL, Starks HE, Cain KC, Uhlmann RF, Pearlman RA (1994).
Measuring preferences for health states worse than death.
Medical Decision Making 14:9-18.

27. Craig BM, Busschbach JJV (2009).
The episodic random utility model unifies worse than death and better than death TTO responses in health state valuation.
Population Health Metrics 7(3):1-10.

28. Cohen J (1978).
Statistical power analysis for the behavioural sciences.
New York: Academic Press.

29. Brazier J, Roberts J (2004).
The estimation of a preference-based index from the SF-12.
Medical Care 42:851-9.

30. Kind P, Hardman G, Macran S (1999).
UK population norms for EQ-5D.
Centre for Health Economics Discussion Paper Series, University of York.

31. Drummond MF, O'Brien BJ, Stoddart GL, Torrance GW (1997).
Methods for the economic evaluation of health care programmes.
Oxford: Oxford Medical Publications.

32. Brazier J, Yang Y, Tsuchiya A, Rowen D (2009).
A review of studies mapping (or cross walking) from non-preference based measures of health to generic preference-based measures.
European Journal of Health Economics (in press).

33. NICE (2004).
Guide to the methods of technology appraisal.
(http://www.nice.org.uk/pdf/TAP_Methods.pdf)

34. Stevens K, Brazier J, McKenna S, Doward L, Cork M (2005).
The development of a preference-based measure of health in children with atopic dermatitis.
British Journal of Dermatology 153:372-7.

Figure captions

Figure 1: Item map for physical functioning, role functioning and pain dimension

Table 1: Characteristics of respondents

	Included respondents (n=344)[1]	*South Yorkshire*[2]	*England*[2]
Mean age (s.d.)	47.8 (18.5)	NA	NA
Age distribution			
18-40	38.6%	41.2%	41.6%
41-65	42.7%	39.1%	39.1%
Over 65	17.2%	19.7%	19.3%
Female	61.9%	51.2%	51.3%
Married/Partner	57.0%	NA	NA
Employed or self-employed	43.6%	56.1%	60.9%
Unemployed	0.6%	4.1%	3.4%
Long-term sick	6.4%	7.7%	5.3%
Full-time student	7.3%	7.5%	7.3%
Retired	24.7%	14.4%	13.5%
Own home outright or with a mortgage	69.2%	64.0%	68.7%
Renting property	29.9%	36.0%	31.3%
Secondary school is highest level of education	41%	NA	NA
EQ-5D score (s.d.)	0.82 (0.26)	NA	0.86 (0.23)
TTO completion rate	98.5%		

[1] Six respondents were excluded; three for valuing all states as identical and less than 1; two for valuing the worst possible state higher than every other state; one for valuing all states as worse than dead.
[2] Statistics for South Yorkshire Health Authority and for England in the Census 2001. Questions used in this study and the census are not identical. The census includes persons aged 16 and above whereas this study only surveys persons aged 18 and above. Age distribution is here reported as the percentage of all adults aged 18 and over.
[3] Interviews conducted in the Measurement and Valuation of Health (MVH) study in 1993 [30].

Table 2: Summary of psychometric and Rasch analysis used to select items per dimension

		Psychometric analysis					*Rasch analysis*						
Items	*Item summary*	*% response at floor (very much)*	*% response at ceiling (not at all)*	*% missing data*	*Correlation with domain score*	*SRM*	*Item range*	*Residual*	*Item level chi-sq P-value*	*Spread at logit zero*	*Disor-dered*	*DIF charac-teristic*	*Poorly fitting item in Rasch model (chi-sq P-value < 0.01)*
Physical and role functioning and pain													
1	Trouble doing strenuous activities	28.5	14.4	0.6	-0.869	-0.206						Sex	
2	Trouble taking a long walk	26.9	16.9	1.1	-0.896	-0.195	-2.576 to 0.516	-1.540	0.385	0.93 to 0.37			
3	Trouble taking a short walk	8.1	46.7	1.7	-0.844	-0.235						Sex	
4	Need to stay in bed or a chair during the day	9.3	36	1.4	-0.788	-0.289	-1.158 to 2.385	0.874	0.120	0.76 to 0.08			
5	Need help with eating, dressing, washing or using the toilet	4.1	72.8	0.6	-0.575	-0.313	1.220 to 2.584	-0.678	0.083	0.23 to 0.07			
6	Limited in doing your work or daily activities	18.5	25.8	1.5	-0.953	-0.259							Yes
7	Limited in pursuing hobbies or other leisure time activities	17.9	30.1	1.2	-0.954	-0.237	-1.526 to -1.192	-0.629	0.600	0.82 to 0.23			
9	Pain	20	20.2	0.9	0.926	-0.499	-2.176 to 1.004	4.685	0.075	0.90 to 0.27			
19	Pain interfered with daily activities	20	30.4	0.5	0.936	-0.432	-1.427 to 0.987	-0.844	0.138	0.81 to 0.27			
Social functioning													
26	Physical condition or medical treatment interfered with family life	8.7	49.2	0.8	-0.899	-0.085	-1.247 to 2.013	1.146	0.175	0.78 to 0.12			
27	Physical condition or medical treatment interfered with social life	12.8	41.8	0.6	-0.942	-0.090	-1.869 to 1.144	0.555	0.345	0.87 to 0.24			

		Psychometric analysis					*Rasch analysis*						
Items	*Item summary*	*% response at floor (very much)*	*% response at ceiling (not at all)*	*% missing data*	*Correlation with domain score*	*SRM*	*Item range*	*Residual*	*Item level chi-sq P-value*	*Spread at logit zero*	*Disor-dered*	*DIF charac-teristic*	*Poorly fitting item in Rasch model (chi-sq P-value < 0.01)*
Emotional functioning													
21	Feel tense	6.6	37.7	0.9	-0.834	-0.264						Age	
22	Worried	12.8	26.0	0.2	-0.860	-0.477						Sex	
23	Feel irritable	4.3	44.3	0.5	-0.776	-0.223	-1.546 to 2.740	3.912	0.442	0.82 to 0.06			
24	Feel depressed	7.8	41.5	0.8	-0.834	-0.330	-1.696 to 2.017	-0.742	0.232	0.85 to 0.12			
Digestion													
13	Lacked appetite	8.9	48.9	0.5	No domain	-0.348	-1.349 to -0.314	-2.244	0.027	0.79 to 0.58			
14	Felt nauseated	2.4	73.9	0.3	No domain	-0.124							Yes
15	Vomited	1.2	87.9	0.3	No domain	-0.046					Yes	Sex	
16	Constipated	7.3	52.7	0.5	No domain	-0.225	-1.258 to -0.449	-1.220	0.231	0.78 to 0.61			
17	Had diarrhoea	0.8	79.5	0.9	No domain	0.058	-0.081 to 1.083	1.029	0.027	0.52 to 0.25			
Fatigue and trouble sleeping													
10	Needed to rest	15.9	14	1.4	No domain	-0.326	-3.113 to 2.212	1.409	0.881	0.96 to 0.10			
11	Trouble sleeping	10.1	41.4	0.2	No domain	-0.276							Yes
12	Felt weak	12.7	22.6	0.6	No domain	-0.296	-2.155 to 2.641	-0.433	0.043	0.90 to 0.07			
18	Tired	12.8	15.9	0.3	No domain	-0.245	-2.933 to 2.805	-0.727	0.540	0.95 to 0.06			

Note: DIF by sex is split male/female and DIF by age is split <65/65+.

Table 3: Goodness of fit for the Rasch model for each dimension

	Item-trait interaction				
Dimension	***Chi-sq (degrees of freedom)***	***P-value***	***Item fit (SD)***	***Person fit (SD)***	***Person separation index***
Physical and role functioning and pain	121.36 (88)	0.01	-0.58(2.12)	-0.32(1.02)	0.90
Social functioning	10.82 (8)	0.21	0.85(0.42)	-0.44(0.80)	0.79
Emotional functioning	65.34 (54)	0.14	-0.18(2.14)	-0.35(0.96)	0.85
Digestion	72.08 (35)	0.00	-0.67(1.20)	-0.30(0.75)	0.47
Fatigue and trouble sleeping	22.05 (20)	0.34	0.08(1.16)	-0.65(1.27)	0.83

Table 4: Summary of EORTC QLQ-30 items included in the EORTC-8D descriptive system

EORTC-8D dimension	*EORTC QLQ-C30 items*	*Question*
Physical functioning	2	Trouble taking a long walk
	3	Extra level added from 'trouble taking a short walk'
Role functioning	7	Limited in pursuing hobbies or other leisure time activities
Pain	19	Pain interfered with daily activities
Social functioning	27	Physical condition or medical treatment interfered with social life
Emotional functioning	24	Felt depressed
Nausea	14	Felt nauseated
Constipation and diarrhoea	16	Constipated
	17	Diarrhoea
Fatigue and trouble sleeping	18	Tired

Table 5: EORTC-8D descriptive system

During the past week:

Physical functioning
You had no trouble taking a long walk
You had a little trouble taking a long walk
You had quite a bit of trouble taking a long walk
You had very much trouble taking a long walk
You had very much trouble taking a short walk outside of the house

Role functioning
You were not limited in pursuing your hobbies or other leisure time activities
You were limited a little in pursuing your hobbies or other leisure time activities
You were limited quite a bit in pursuing your hobbies or other leisure time activities
You were limited very much in pursuing your hobbies or other leisure time activities

Pain
Pain did not interfere with your daily activities
Pain interfered a little with your daily activities
Pain interfered quite a bit with your daily activities
Pain interfered very much with your daily activities

Emotional functioning
You did not feel depressed
You felt a little depressed
You felt quite a bit depressed
You felt depressed very much

Social functioning
Your physical condition or medical treatment did not interfere with your social activities
Your physical condition or medical treatment interfered a little with your social activities
Your physical condition or medical treatment interfered quite a bit with your social activities
Your physical condition or medical treatment interfered very much with your social activities

Fatigue and sleep disturbance
You were not tired
You were a little tired
You were quite a bit tired
You were tired very much

Nausea
You did not feel nauseated
You felt a little nauseated
You felt nauseated quite a bit
You felt nauseated very much

Constipation and diarrhoea
You were not constipated and did not have diarrhoea
You were constipated and/or had diarrhoea a little
You were constipated and/or had diarrhoea quite a bit
You were constipated and/or had diarrhoea very much

Table 6: Estimated preference weights

Dimensions and levels	*(1) OLS*	*(2) MLE RE*	*(3) ERUM OLS*	*(4) Mean model*		*(5) Consistent mean model*
PF2	**0.061**	**0.052**	**0.052**	**0.065**	PF2	**0.065**
PF3	**0.076**	**0.079**	**0.077**	**0.078**	PF3	**0.078**
PF4	**0.135**	**0.134**	**0.103**	**0.139**	PF45	**0.127**
PF5	**0.121**	**0.127**	**0.104**	**0.105**		
RF2	0.026	0.023	**0.044**	0.032	RF2	0.032
RF3	**0.042**	**0.052**	**0.050**	**0.045**	RF3	**0.045**
RF4	**0.082**	**0.090**	**0.076**	**0.079**	RF4	**0.078**
PAIN2	**0.059**	**0.041**	**0.054**	**0.059**	PAIN2	**0.059**
PAIN3	**0.060**	**0.060**	**0.064**	**0.062**	PAIN3	**0.062**
PAIN4	**0.070**	**0.083**	**0.070**	**0.065**	PAIN4	**0.064**
EF2	0.028	0.027	**0.032**	0.030	EF2	0.030
EF3	**0.063**	**0.072**	**0.053**	**0.066**	EF3	**0.066**
EF4	**0.157**	**0.160**	**0.132**	**0.150**	EF4	**0.149**
SF2	0.025	0.022	0.029	0.027	SF2	0.027
SF3	**0.059**	**0.065**	**0.046**	**0.059**	SF3	**0.059**
SF4	**0.173**	**0.174**	**0.132**	**0.163**	SF4	**0.163**
FAT2	**0.046**	0.026	**0.038**	**0.046**	FAT2	**0.047**
FAT3	**0.052**	**0.031**	**0.052**	**0.054**	FAT3	**0.054**
FAT4	**0.104**	**0.064**	**0.084**	**0.093**	FAT4	**0.092**
NAU2	**0.031**	**0.036**	0.025	0.032	NAU23	0.026
NAU3	0.015	**0.037**	**0.027**	0.019		
NAU4	**0.062**	**0.079**	**0.052**	**0.057**	NAU4	**0.056**
CD2	0.012	0.022	0.011	0.016	CD2	0.016
CD3	**0.050**	**0.037**	**0.035**	**0.052**	CD3	**0.052**
CD4	**0.078**	**0.070**	**0.059**	**0.073**	CD4	**0.072**
Observations	2710	2710	2710	85		85
R-squared	0.60		0.56	0.97		0.97
Number of id		344				
Inconsistencies	2	1	0	2		0
Insignificant level coefficients	5	5	3	6		5
MAE	0.052	0.054	0.046	0.050		0.051
MAE>0.05	41	41	33	37		39
MAE>-0.10	9	13	6	8		9

Note: Figures in bold have t-statistics significant at the 5% level
PF=Physical functioning, RF=Role functioning, PAIN=Pain, EF=Emotional functioning, SF=Social functioning, FAT=Fatigue, NAU=Nausea, CD=Constipation and/or diarrhoea. PF23=Physical functioning at level 2 or 3, NAU23=Nausea at level 2 or 3.

Figure 1: Item map for physical functioning, role functioning and pain dimension