Deriving a preference-based measure for cancer using the EORTC QLQ-C30: a confirmatory versus exploratory approach

Authors: Daniel S. J. Costa, Madeleine T. King, Neil K. Aaronson, Peter M. Fayers, Peter S. Grimison, Monika Janda, Julie F. Pallant, Donna Rowen, Galina Velikova, Rosalie Viney, Tracey A. Young; and the MAUCa consortium.

ABSTRACT

Background: To derive preference-based measures from various condition-specific descriptive health-related quality of life (HRQOL) measures. A general 2-stage method is evolved: 1) an item from each domain of the HRQOL measure is selected to form a health state classification system (HSCS); 2) a sample of health states is valued and an algorithm derived for estimating the utility of all possible health states. The aim of this analysis was to determine whether confirmatory or exploratory factor analysis (CFA, EFA) should be used to derive a cancer-specific utility measure from the EORTC QLQ-C30.

Methods: Data were collected with the QLQ-C30v3 from 356 patients receiving palliative radiotherapy for recurrent or metastatic cancer (various primary sites). The dimensional structure of the QLQ-C30 was tested with EFA and CFA, the latter based on a conceptual model (the established domain structure of the QLQ-C30: physical, role, emotional, social and cognitive functioning, plus several symptoms) and clinical considerations (views of both patients and clinicians about issues relevant to HRQOL in cancer). The dimensions determined by each method were then subjected to item response theory, including Rasch analysis.

Results: CFA results generally supported the proposed conceptual model, with residual correlations requiring only minor adjustments (namely, introduction of two cross-loadings) to improve model fit (increment $\chi^2(2) = 77.78$, $p < .001$). Although EFA revealed a structure similar to the CFA, some items had loadings that were difficult to interpret. Further assessment of dimensionality with Rasch analysis aligned the EFA dimensions more closely with the CFA dimensions. Three items exhibited floor effects (>75% observation at lowest score), 6 exhibited misfit to the Rasch model (fit residual > 2.5), none exhibited disordered item response thresholds, 4 exhibited DIF by gender or cancer site. Upon inspection of the remaining items, three were considered relatively less clinically important than the remaining nine.

Conclusions: CFA appears more appropriate than EFA, given the well-established structure of the QLQ-C30 and its clinical relevance. Further, the confirmatory approach produced more interpretable results than the exploratory approach. Other aspects of the general method remain largely the same. The revised method will be applied to a large number of data sets as part of the international and interdisciplinary project to develop a multi-attribute utility instrument for cancer (MAUCa).

Keywords: cancer, EORTC QLQ-C30, multi-dimensional health state classification system, QALYs.

INTRODUCTION

In making decisions about which cancer treatments provide the most value for patients in terms of clinical benefit, there is increasing emphasis on health-related quality of life (HRQOL) and financial cost as well as survival. For reimbursement decisions, many jurisdictions either now require or prefer cost-utility analysis (CUA) [1, 2]. CUA is a type of cost-effectiveness analysis in which the measure of health outcome combines effects on both HRQOL and survival in a single metric, most often the Quality Adjusted Life Year (QALY). The key metric for such analyses is 'utility', a preference-based index of HRQOL in which 1 is equivalent to full health and 0 is equivalent to being dead [3, 4].

Standard questionnaires for assessing health status and HRQOL, such as the generic SF-36 [5] and the cancer-specific QLQ-C30 [6], provide a standardised description of a wide range of health states in terms of distinct levels of function or symptom experience. A scoring algorithm provides summary scores, typically one per domain, and sometimes one or two overall scores [7, 8]. These scores are not an appropriate metric to estimate utilities and QALYs because they may not reflect respondents' strength of preference among dimensions of HRQOL and versus survival. Obtaining utilities requires the use of preference-based assessment. The most commonly used methods are the standard gamble or time trade-off task [3] , although other choice-based tasks have been explored, such as discrete choice experiments [9]. Alternatively, multi-attribute utility instruments (MAUI) combine self-report of HRQOL with a preference based scoring algorithm. Each unique combination of item levels and dimensions is called a health state. There are typically a very large number of health states. For example, the EQ-5D has 243 health states, being the 3^5 unique combinations of 5 dimensions, each with 3 levels [10]. The scoring algorithm is based on valuations, typically by a general population sample of a sample of health states, derived using one of the preference-based tasks described above. Statistical analysis of these valuations yields a utility weight for each health state covered by the health state classification system

of the instrument. Once the valuation is conducted and the scoring algorithm derived, the latter can then be applied to HRQOL data collected prospectively with the descriptive component of the MAUI. Arguably, valuations may differ by country [11], and country-specific utility weights may need to be generated.

A MAUI can be used in clinical populations to assess the impacts of disease or treatment on HRQOL (using the descriptive health state classification component), the responses can be converted to utilities (with the scoring algorithm), and thence used to value differences in health status using the QALY metric. Generic MAUIS include the EQ-5D [10], Health Utilities Index [12]; HUI3 [13], and SF-6D[14]; and are used for comparison across a range of health conditions. Condition-specific MAUIs may be more sensitive than generic MAUIs to detect the key health impacts of particular diseases and treatments, and are increasingly common [15-17].

The European Organisation for Research and Treatment of Cancer's (EORTC) core Quality of Life Questionnaire, the QLQ-C30 [6], is one of the most widely used cancer-specific HRQOL instruments. Its wide use in cancer clinical trials makes it an ideal candidate for adaptation to a condition-specific MAUI that would allow the conduct of cost-utility analysis without the need for additional data collection. In order to make the health states described by the QLQ-C30 amenable to valuation using preference-based techniques [3], the number of items must be reduced so that the number of health states is not prohibitively large.

Early in its development, the general process of reducing descriptive HRQOL instruments to form health state classification systems was guided by expert judgement [14, 18]. A more recently employed strategy is to also consider statistical and psychometric properties of items [19]. Specifically, exploratory

factor analysis (EFA) has been used to identify structurally independent dimensions (a theoretical requirement for MAUIs[20]), followed by Rasch analysis to assess psychometric properties of items relevant to their performance in a MAUI [19, 21]. Items are first excluded on the basis of various statistical and psychometric criteria related to both EFA and Rasch procedures, and then one item (or a few) from each domain are finally selected for inclusion, with the general aim of reducing the number of items without substantial loss of information. The main advantages of this method are that the resulting classification system represents the dimensionality of the measure using observed data, and further this method can be used for all measures regardless of how well established the dimensional structure is. One crucial disadvantage is that EFA will produce only factors, and to move beyond this to establishing coherent dimensions the results need to be interpreted and clinical advice is required, but the method in which this is done is subjective.

In the case where health state classification systems are to be derived from a validated questionnaire with a well-established dimensional structure, such as the QLQ-C30, arguably a confirmatory approach to the question of dimensionality is more appropriate than an exploratory approach. The confirmatory approach involves the positing of a specific dimensional structure (the conceptual model) that is tested with CFA). This has three advantages over the exploratory approach. First, many of the arbitrary decisions involved in EFA (e.g., method of extraction, method of rotation, number of factors to extract) are removed, replaced instead with more theoretically or clinically driven decisions, such as which items are hypothesised to load on which factors. Second, without a priori theoretical or practical guidance any given solution may be difficult to interpret. Third, the positing of a specific model allows clinical considerations – which we define here as the views of both patients and clinicians about issues relevant to HRQOL in cancer – to play a more structured a priori role than EFA can allow. Certain items may be included in or excluded from the model a priori, based on clinical or theoretical considerations. Such a

priori decisions can also be made in relation to whether items can be directly included in the health state classification system without being included in CFA. Although this can also be done using EFA, in the confirmatory approach clinical considerations are incorporated into the item *assessment* procedure, rather than the *selection* of specific items. An advantage of the confirmatory approach is that clinical considerations can be built in to the general method of item assessment, rather than acting as a post-hoc, context-specific activity. Items deemed important in the trade off between HRQOL and survival may thus be selected solely according to clinical considerations. For such items, clinical considerations would override statistical criteria, ensuring that the condition-specific preference-based measure contains symptoms of particular relevance to that condition. In cancer, these include fatigue, pain and nausea[22, 23].

The analysis reported in this paper is part of a large-scale international, interdisciplinary collaboration whose primary aim is to develop cancer-specific multi-attribute utility instruments (MAUCa) from the QLQ-C30 and the other most widely used cancer-specific HRQOL instrument, the FACT-G. In an earlier study, the methods of Young et al [21, 24], including EFA, were applied to develop a preference-based measure from the QLQ-C30 using a UK sample of multiple myeloma patients [25]. The aim of the current paper is to build on this previous work [21, 25] by assessing whether the confirmatory approach proposed above is more appropriate and useful than an exploratory approach to derive a cancer-specific MAUI from the QLQ-C30, given its well-established domain structure. Note that the objective of this paper was not to develop a specific health state classification system, but rather to refine and make further recommendations on the appropriate methodology for defining the dimension structure for the MAUI, focusing on Step 1 of the 7 step item selection procedure described by Young et al [19].

Methods

Ethical approval for this study was granted by the University of Sydney Human Research Ethics Committee (Protocol No.: 13207).

Quality of life instrument

The QLQ-C30 (version 3) is a multidimensional instrument containing 30 items assessing symptoms, functioning and overall HRQOL (Table 1). Its validity and reliability as a self-report measure of the impact of cancer and its treatment on HRQOL are well established [6, 26]. Responses to items 1-28 are made on a four-point scale (1 = "Not at all", 2 = "A little", 3 = "Quite a bit", 4 = "Very much"), and responses to items 29 and 30 (global health and quality of life items) are made on a seven-point scale (1 = "Very poor" and 7 = "Excellent"). Respondents are instructed to assess items 6 to 30 as experienced during the past week. The 30 items form five functioning scales, three multi-item symptom scales, five single-item symptom scales (plus a financial difficulties item) and a global health status and QoL scale (Table 1).

[Table 1 here]

Data set

A secondary analysis was conducted on data collected with the QLQ-C30 (version 3) from a sample of 356 patients (53% Norwegian, 47% Swedish) with Stage IV/recurrent/metastatic cancer from a variety of primary sites (36% prostate, 30% breast, 11% lung, 23% other), all undergoing palliative radiotherapy in a randomised clinical trial comparing two fractionations [27]. The mean age was 66.77 years (SD = 10.60, range 31.59 – 90.32) and 43.8% were female.

Analysis

1. Descriptive statistics

Frequencies of response categories for each item were examined to determine if items exhibited floor or ceiling effects (i.e. a preponderance of scores at the minimum and maximum scale score). Functioning items with ceiling effects (i.e. a majority of patients endorsed the highest level of functioning) and symptom scales with floor effects (i.e. a majority of patients endorsed the lowest level of symptoms) make these items poor candidates for the health state classification system for two reasons: 1) they are not common problems and therefore unlikely to affect patient quality of life; and 2) they cannot register improvements and are therefore unlikely to be responsive to beneficial treatments. Conversely, functioning items with floor effects (i.e. a majority of patients endorsed the lowest level of functioning) and symptom scales with ceiling effects (i.e. a majority of patients endorsed the highest level of symptoms) reflect common problems that may therefore have high impact on patient quality of life but are unlikely to be responsive to deterioration.

2. Exploratory versus Confirmatory Factor Analysis

EFA is a statistical procedure in which variables (in this case, items of the QLQ-C30) are grouped into relatively independent subsets based on their inter-correlations, without any prior assumptions about the composition of these subsets. In contrast, CFA involves testing a pre-specified arrangement of items into subsets, guided by a conceptual model. EFA and CFA were conducted to assess the dimensional structure of the QLQ-C30 (that is, how items cluster into dimensions), and the results compared. The model of HRQOL that we tested with CFA was based on both the established structure of the QLQ-C30 [28] and clinical considerations (described below). Judgements about the meaningfulness of the item clusters obtained using EFA, i.e., whether the items are conceptually related, were made.

Three items were excluded a priori from both the EFA and CFA. Item 28 (financial difficulties) was excluded from all analyses as it is neither a symptom nor a measure of functioning. The two global items (29 and 30) were also excluded because each item in the health state classification system should represent a specific domain of HRQOL (functioning or symptom) rather than global quality of life [16].

2.1. Exploratory approach

Certain technical parameters need to be specified in EFA. For our primary EFA, we used the principal axis factoring (PAF) method to extract factors, with a direct oblimin rotation to allow factors to be correlated. We examined the suitability of the data for EFA (i.e., the degree of correlation amongst the items) with the Kaiser-Myer-Olkin (KMO) measure of sampling adequacy and the Bartlett test of sphericity. Criteria for suitability are KMO > .8 and a *p*-value for Bartlett's χ^2 of less than .01 [29]. We used parallel analysis [30], implemented in specialist software (Monte Carlo PCA for Parallel Analysis), to inform selection of factors. This involves computing mean eigenvalues from randomly-generated sets of data (1000 were used in this case) of the same size (number of items and number of observations) as the observed data set. Any factor obtained from the observed data set with an eigenvalue exceeding the corresponding eigenvalue generated from parallel analysis was considered for selection. A scree plot was also inspected. An item was considered to load on a factor if it had a factor loading in the pattern matrix greater than .3, and did not load on any other component.

We also conducted a sensitivity analysis involving all 15 combinations of: three extraction methods (PAF, maximum likelihood (ML), principal components (PC)) and five rotation methods (oblimin, promax, varimax, equamax, quartimax), comparing the degree of variability in solutions obtained due to variation in these technical parameters.

2.2. Confirmatory approach

A priori clinical considerations

The guiding principle here was to consider which aspects of functioning, symptoms and side-effects should be included in the health state classification system, and hence the utility function of cancer-specific MAUI, in order for it to have face validity for economic evaluation of cancer treatments. Inclusion of dimensions was determined by three considerations: first, the dimensions available in the QLQ-C30; second, the patient's perspective (which symptoms, side-effects and aspects of functioning are considered important by patients in their overall assessment of quality of life); third, the clinician's perspective (which dimensions matter when assessing the value of alternative treatments). Previous research has shown that patients [22] and clinicians [23] consider pain, fatigue, nausea/vomiting, constipation and diarrhoea to be important. All are available in the QLQ-C30. It is also well-established that the various aspects of functioning are correlated with measures of overall quality of life [28]. Regression analysis has also revealed certain domains to be strong predictors of global quality of life, e.g., emotional functioning and fatigue [31].

The primary difference between clinical considerations using the confirmatory approach and clinical considerations as they have been employed in previous exploratory approaches is that in the confirmatory approach clinical considerations are incorporated a priori as part of the procedure to assess the items for inclusion, and are thus allocated a more structured role.

Established structure of the QLQ-C30

We defined the *conceptual model* as the arrangement of items on the QLQ-C30 into domains based on the established structure of the QLQ-C30 [6] and the clinical considerations described above. We

defined the *measurement model* as the subset of the conceptual model that was empirically tested using CFA.

The conceptual model to be used as a starting point for the QLQ-C30 was thus comprised of the following eight latent variables (with the item numbers of their manifest variables given in parentheses) and five single item domains:

Functioning: physical functioning (items 1-5); role functioning (items 6-7); emotional functioning (items 21-24); social functioning (items 26-27); cognitive functioning (items 20 and 25).

Symptoms: pain (items 9 and 19); fatigue (items 10, 12 and 18); nausea and vomiting (items 14-15); dyspnoea (item 8); sleep (item 11); appetite (item 13), constipation (item 16) and diarrhoea (item 17).

Items included *a priori* in the conceptual model and therefore excluded from measurement model: Dyspnoea, sleep, appetite, constipation and diarrhoea were considered of sufficient clinical importance for consideration in the health state classification system, but as these domains are represented by single items (8, 11, 13, 16 and 17 respectively), these items were excluded from the measurement model.

CFA based on the conceptual models described above was run using AMOS version 19. In CFA, an estimation method, parallel to the extraction method used in EFA, must be specified, but it is only in rare circumstances that an estimation method other than maximum likelihood is used [32]. A rotation method does not need to be specified in CFA. Correlations amongst the latent variables were not constrained, while correlations between error terms were fixed to 0. The fit of the model to the data was assessed using the following indices and their corresponding widely-accepted cut-off guidelines indicating good model fit, [32]: chi-squared statistic / degrees of freedom (less than 2); comparative fit

index (CFI, greater than .9); goodness-of-fit index (GFI, greater than .9); root mean square error of approximation (RMSEA, less than .05). Additionally, the residual correlation matrix was examined – any residual correlation greater than .1 was further examined [33]. If model fit was poor on any one of the measures, then factor loadings and residual correlations were examined in order to determine alterations to the model that improve fit. The model was modified and re-tested until a model was obtained that was conceptually meaningful and also adequately fitted the data.

3. Item assessment

The assessment of items for selection in the health state classification system, done separately for the factor solutions obtained from EFA and CFA, was subject to a number of criteria, informed by a variety of techniques, described below. These techniques have already been described in detail by Young *et al* and interested readers are referred to step 2 of their guidance for deriving a MAUI, the details of this step are summarised briefly below. Although the purpose of this paper is not to generate a health state classification system, the items were nonetheless assessed against these criteria, with a view to facilitating decisions about item selection.

2a. Rasch analysis

The primary technique used was Rasch analysis, in which the observed responses to items are assumed to reflect an underlying latent variable, such that the probability of endorsing an item is a monotonic increasing function of the underlying latent variable. Items that met criteria described below were deemed to conform to the Rasch model [34] and were therefore retained for consideration in the health state classification system.

All Rasch analysis was conducted using RUMM2020 and was performed separately for the dimensions identified using both types of factor analysis. The initial stage of Rasch analysis was conducted with the aim of determining whether any of the items exhibited problems with fit to the model, item response threshold ordering or differential item functioning (DIF) [35]. Local dependence was also assessed, with any items exhibiting redundancy considered for exclusion. Any items that exhibited such problems were considered for exclusion from the health state classification system, although not from the Rasch analysis (except where indicated).

(a) Poor item fit

The overall fit of the Rasch model was examined using the item-trait interaction chi-squared statistic. Good model fit was considered to be reflected by a non-significant chi-squared statistic. A Bonferroni correction was applied to the criterion of significance, such that the adjusted critical *p* value was .05 divided by the number of items. Generally, misfit can be classified as either due to persons or items by examining the standard deviation of the fit residual for each – the fit statistics are transformed to approximate standard normal distribution, so values for the standard deviation that vary greatly from 1 indicate misfit; we took values greater than 1.5 to indicate the presence of misfitting items. Poor fit due to either source was assessed further by examining fit residuals for individual items or persons – items with fit residuals that exceeded 2.5 were removed from the Rasch analysis, and persons with fit residuals that exceeded 2.5 were removed only if they appeared to contribute to item misfit. This process was repeated until only well-fitting items remained and the overall goodness of fit of the model was non-significant. Any items excluded due to misfit were kept aside and assessed according to other criteria, including descriptive statistics and clinical considerations (described below).

(b) Assessment of response format

An appropriately functioning item requires a response format that respondents use in a consistent manner. Examining response thresholds – the points at which each consecutive response category for an item is equally likely to be endorsed – allows the assessment of response format in this regard. For an appropriately functioning (i.e., ordered) item, the response thresholds between successive categories should be ordered, such that the threshold between categories 1 and 2 falls below the threshold between categories 2 and 3, and so on. A disordered response threshold indicates that respondents are not able to adequately discriminate between the response categories they are asked to select from. Item-threshold probability curves were examined for all items to determine whether there was any disordering in the response thresholds.

(c) Invariance of item functioning across different groups

For inclusion in the health state classification system, an item should function in the same way across different groups (e.g., gender). Specifically, the probability of selecting a certain response category for a given value of the latent trait should be invariant across groups. If it is not, the item exhibits differential item function (DIF). DIF is a form of bias in which systematic differences in patterns of responding to an item are observed between individuals with different characteristics, despite having the same level of the latent variable (reflected by their logit score). If a group shows a consistent difference in item responses to another group across the range of values for the latent variable, this is known as *uniform DIF*. *Non-uniform DIF* occurs when the differences between groups vary over the range of values of the latent variable.

The data were examined for DIF across gender and cancer site. (DIF across country is an important issue but has been examined previously [36]). Because cross-population comparisons using the health state

classification system are desirable, any items exhibiting DIF was excluded from the health state classification system.

(d) Local dependence

Although inter-item correlations were examined prior to Rasch analysis, they were further examined using residual principal components analysis in RUMM2020. The primary purpose of this was to determine whether any items within a domain formed a 'subtest' (i.e., a single item comprised of the information contained in multiple correlated items). Any pair of items with residual correlation that was at least .3 above the mean residual correlation for all pairs of items was considered to be locally dependent, and the possibility of combining these items into a single item covering the content of both items in the pair for inclusion in the health state classification system was considered.

Results

Descriptive statistics

The frequencies of the response categories for each of the 27 items included in the analysis are presented in Figure 1. For 3 items, over 75% of respondents reported having no problem at all: items 5 (need help eating, dressing, washing), 15 (vomiting) and 17 (diarrhoea). These figures are consistent with extensive reference data for the QLQ-C30 [37]. There were no other problems with floor or ceiling effects.

Exploratory approach

Table 2a provides a summary of the results from the primary EFA (PAF extraction and oblimin rotation) and related Rasch analyses. The inter-item correlations were adequate for factor analysis (KMO = .892; Bartlett's χ^2 = 3993.58, $p < .0005$). Parallel analysis suggested the extraction of three factors, and this was supported by inspection of the scree plot. Items 8 (dyspnoea), 16 (constipation), 17 (diarrhoea) and 25 (memory) loaded weakly on all factors[1], while cross-loadings were observed for items 12 and 18 (both fatigue items).

[Table 2a here]

The three factors (denoted EFA Factors 1, 2 and 3 for clarity of presentation) determined for Rasch analysis were:

EFA Factor 1: Items 1-7, 9, 10, 19 and 27 (encompassing the physical and role functioning domains, the two pain items, one of the three fatigue items and one of the two social functioning items);

EFA Factor 2: Item 11, 20-24 and 26 (encompassing the emotional functioning domain, the insomnia item, one of the two cognitive functioning items and one of the two social functioning items); and

EFA Factor 3: Items 12-15, 18 (encompassing two of the three fatigue items, the appetite loss item, the two nausea/vomiting items. The two cross-loading factors (fatigue) were assigned to this factor because they are symptoms that are more closely related to the items on this factor than Factor 2.

The results of EFA differed slightly depending on the extraction and rotation method used. Using all 15 combinations of methods: items 1-7, 9 and 19 loaded on Factor 1; items 11, 21-24 and 26 loaded on Factor 2; items 13-15 loaded on Factor 3, and; items 8 and 16 exhibited weak loadings on all factors.

There were a few noteworthy differences. Items 17 (diarrhea, Factor 3) and 25 (memory, Factor 2/Factor3) had stronger loadings for PCA than for PAF and ML, to the extent that, using a loading cut-off of 0.3, they would have been comfortably included in the PCA solution, but not PAF or ML. For items 12 (weak) and 18 (tired), for all extraction methods loadings were strongest for Factors 2 and 3 except for when quartimax rotation was used; in this case, Factor 1 exhibited the dominant loadings. For items 10 (rest) and 27 (interfered with social activities) Factor 1 exhibited the dominant loading but strength of cross-loadings differed between extraction/rotation combinations, and the same for item 20 (concentration) except Factor 2 dominated. Results are available from authors on request.

Confirmatory approach

The factor loadings obtained from CFA are presented in Table 2b. The loadings of all items on their respective factors were relatively strong and all statistically significant ($p < .001$). Model fit was adequate ($\chi^2/df = 2.04$, CFI = .945, GFI = .907, RMSEA = .058). Residual correlations suggested additional relations between items 4 and 10, items 9 and 10, items 24 and 26, and items 15 and 24. Items 4 and 10 had a high residual correlation (.23) and cover similar content (needing to rest), so cross-loadings were introduced for item 4 (with Fatigue) and item 10 (with Physical Functioning). These cross-loadings resulted in improved model fit ($\chi^2/df = 1.63$, CFI = .967, GFI = .924, RMSEA = .045). The other pairs mentioned do not cover similar content and had smaller residual correlations (highest .14) and so no modifications were made to the model on the basis of their correlations.

The correlations between the eight factors are displayed in Table 3. Most noteworthy was the very high (.855) correlation between role and physical functioning, suggesting that the items in these two factors may reflect a single factor.

Although the hypothesised eight-factor structure of the QLQ-C30 was generally supported, it was decided that the physical functioning domain (items 1-5) be combined with the role functioning domain (items 6 and 7) as well as item 10 for the purpose of Rasch analysis, based on the results above. Item 10 was not included in the fatigue domain (with items 12 and 18) for Rasch analysis. The other domains were subjected to Rasch analysis without any change from the factor specified a priori.

Rasch analysis

(1) Based on EFA

Factor 1: Items 1-7, 9, 10, 19 and 27

The overall item fit was poor, with a fit residual standard deviation (FRSD) of 2.44. Person fit was good (FRSD = 1.043) and the person-item interaction χ^2 was highly significant, indicating poor fit to the Rasch model. Individual items 9 (pain, fit residual = 4.96) and 27 (social activities, fit residual = 3.15) exhibited misfit. When these two items were subsequently removed from the analysis, overall item fit was still poor, although no individual items exhibited misfit. Following the removal of these items, high residual correlations were observed between items 2 (long walk) and 3 (short walk), items 4 (stay in bed) and 10 (need to rest), and items 6 (daily activities) and 7 (leisure activities). The correlation between items 6 and 7 was unsurprising, as the traditional QLQ-C30 domain structure treats these as a separate domain to the other items. The other two pairs of residual correlations are also unsurprising, given the content of the items. These residual correlations may explain the poor overall item fit. No item thresholds were disordered. Item 1 (strenuous activities) exhibited uniform DIF for both gender and cancer site. Specifically, for a constant level of the latent trait, females scored consistently higher than males on this item, and those with prostate cancer scored consistently lower on this item than the other groups.

Factor 2: Items 11, 20-24, 26

The overall item fit was poor (FRSD = 2.82). Person fit was good (FRSD = 1.11) and the person-item interaction χ^2 was highly significant, indicating poor fit to the Rasch model. Individual items 11 (insomnia, fit residual = 4.56) and 26 (family life, fit residual = 2.79) exhibited misfit. When these two items were subsequently removed from the analysis, overall item fit was still poor (FRSD = 2.34), with item 20 (concentrating) exhibiting misfit (fit residual = 3.18). When item 20 was removed, item 23 (irritable) then exhibited misfit (fit residual 3.29), and overall item fit was finally achieved (FRSD = 1.41) when item 23 was removed. Following the removal of these items, no residual correlations were observed and no item thresholds were disordered. Items 21 (tense) and 22 (worry) exhibited uniform DIF for gender, with females generally scoring higher on these items than males when the latent trait was held constant. Item 22 also exhibited uniform DIF for cancer site, with breast and lung cancer patients generally scoring higher on this item than prostate and other cancer.

Factor 3: Items 12-15, 18

The overall item fit was good (FRSD = 0.848) as was person fit (FRSD = 0.785) and the person-item interaction (p = .3). No individual items exhibited misfit. Large residual correlations were observed between items 12 and 18, and items 14 and 15. Neither correlation was surprising, as items 12 and 18 are both fatigue items and items 14 and 15 are both nausea/vomiting items. No item thresholds were disordered. Item 14 (nausea) exhibited uniform DIF for cancer site, with breast cancer patients scoring consistently lower on this item than other cancer patients of the same level of the latent trait.

2. Based on CFA

Table 2b provides a summary of the results from the CFA and related Rasch analyses.

Factor 1: Items 1-7, 10

Person fit was good (FRSD = 0.848) and the person-item interaction χ^2 was not significant (p = .051), indicating adequate fit to the Rasch model. Although overall item fit residual standard deviation exceeded the recommended guidelines (FRSD = 1.60), no individual items exhibited misfit. Large residual correlations were found between items 2 and 3, items 4 and 10, and items 6 and 7. No item thresholds were disordered. Item 1 exhibited uniform DIF for both gender and cancer site. Specifically, for a constant level of the latent trait, females scored consistently higher than males on this item, and those with prostate cancer scored consistently lower on this item than the other groups.

Factor 2: Items 21-24

The overall item fit was poor (FRSD = 2.28). Person fit was good (FRSD = 1.13), however the person-item interaction χ^2 was highly significant, indicating the presence of misfitting items. Item 23 (fit residual = 3.29) exhibited misfit. Upon removal of item 23, this domain was identical to EFA Factor 2, and so its properties in relation to thresholds, DIF and residual correlation are identical to those described for EFA Factor 2.

Factor 3: Items 26, 27

The overall item (FRSD = 0.23) and person fit were good (FRSD = 0.77) and the person-item interaction χ^2 was not significant (p = .01), indicating adequate fit to the Rasch model. Neither item exhibited misfit or disordered thresholds. Item 27 exhibited uniform DIF by cancer site, such that prostate cancer patients tended to score lower on this item than other patients.

Factor 4: Items 20, 25

The overall item (FRSD = 0.44) and person fit (FRSD = 1.08) were good, and the person-item interaction χ^2 was not significant (p = .07), indicating adequate fit to the Rasch model. Neither item exhibited misfit, disordered thresholds or DIF.

Factor 5: Items 9, 19

The overall item (FRSD = 0.83) and person fit (FRSD = 0.97) wereere good, and the person-item interaction χ^2 was not significant (p = .62), indicating adequate fit to the Rasch model. Neither item exhibited misfit, disordered thresholds or DIF.

Factor 6: Items 12, 18

The overall item (FRSD = 0.32) and person (FRSD = 0.99) were good, and the person-item interaction χ^2 was not significant (p = .80), indicating adequate fit to the Rasch model. Neither item exhibited misfit or disordered thresholds. Item 12 (felt weak) exhibited non-uniform DIF by site, such that breast and prostate cancer patients scored higher on this item than other cancer patients for low values of the latent trait, but lower for higher values of the latent trait.

Factor 7: Items 14, 15

The overall item (FRSD = 0.36) and person (FRSD = 0.85) were good, however the person-item interaction χ^2 was significant (p = .006). Neither item exhibited misfit or disordered thresholds. Both items exhibited non-uniform DIF by site – for item 14, patients with lung cancer scored higher on this item than other patients at high levels of the latent trait, and for item 15, patients with lung cancer scored lower on this item than other patients at high levels of the latent trait.

Discussion

The factor structures obtained from EFA and CFA were similar, and CFA was demonstrated to be a more structured process and to produce more readily interpretable solutions than EFA. Many of the discrepancies between the hypothesised factor structure in CFA and the clusters of items that emerged from EFA were eliminated when the factors obtained from EFA were subjected to Rasch analysis. For example, EFA Factor 2 originally comprised items 11, 20-24 and 26, but following Rasch analysis items 11, 20 and 26 were removed due to misfit, reducing this dimension to the emotional functioning domain of the QLQ-C30 (items 21-24). Item 23 was then further found to misfit and removed. The important point is that the confirmatory approach arrived at this solution more efficiently than the exploratory approach. Furthermore, the one adjustment to the measurement model tested in CFA that was required (namely, the correlation between items 4 and 10) was readily identified and accommodated in the model.

The EFA results were found to differ depending on the method of extraction and rotation employed. Although these differences were not large, they may have had some impact on the item selection process. For example, the inclusion or exclusion of item 17 (diarrhoea) and different decisions about which domain should include the fatigue items (12 and 18) may affect the composition of the health state classification system.

Some aspects of the EFA solution were difficult to interpret. For example, the social functioning items loaded on different factors; specifically, item 26 (interfered with family life) loaded with physical/role functioning items, and item 27 (interfered with social activities) loaded with emotional functioning items. Similarly, fatigue items loaded with nausea, vomiting and lack of appetite. Although post hoc

explanations of these relations are possible (and may well be causal – see below), it is difficult to justify the inclusion of such items in the same domain for the purpose of selecting items for a utility instrument. For example, whether respondents experience interference with social activities is arguably a substantively different issue to whether respondents feel tense, and it seems inappropriate for these two items to be competing candidates for inclusion to represent the same factor in the HSCS. This means that judgement must be applied when using EFA as the factor analysis will establish *factors*, and clinical input and interpretation is required to derive the *dimensions* from these factors. In contrast, in the CFA approach this guidance is provided at the outset to inform the factor analysis, meaning that the results directly represent the dimensionality of the measure. It is worth noting that three of the four items with weak EFA loadings (items 8, 16 and 17) were also three of the five items (along with items 11 and 13) that were excluded from the measurement model a priori, partly on the basis of anticipated poor performance in factor analysis.

EFA produced a solution that combined the physical (items 1-5) and role functioning domains (items 6 & 7) of the QLQ-C30. In the CFA, model fit was adequate with these two domains kept separate, although the two domains were very highly correlated. Residual principal components analysis, as part of the Rasch analysis confirmed that these are in fact two separate domains. One possible reason for this is that items 6 and 7 differ from items 1-5 in their "item difficulty", a phenomenon that would be more readily identified by Rasch analysis than factor analysis. An alternative explanation is that there exists a higher order factor that encompasses both physical and role functioning, or that there is some causal relation between these two factors. These latter possibilities are addressed further below, but are in any case more readily addressed using a confirmatory than an exploratory approach.

The confirmatory approach employed in the present analysis provided a structured role for clinical considerations and an explicitly articulated relation to the statistical and psychometric criteria used in the item selection process, whereas in the previously employed exploratory approach, clinical considerations were less formally specified and explicitly integrated with the statistical analysis. Rowen et al [25] in the derivation of EORTC-8D employed the input of a clinician to ensure the statistical results made sense clinically. In the present analysis we have developed the structured integration of clinical considerations further into the predefined set of judgement criteria. Furthermore, by identifying certain items as of interest a priori allows a structured approach to the selection of items that are of clinical relevance but may not perform adequately in the statistical analysis. For example, although few respondents in this data set reported problems with diarrhoea (item 17), the a priori inclusion of this item in the conceptual model allowed clinical considerations to override the statistical criteria. The importance of this is illustrated by the ALTTO trial, in which diarrhoea was a critical side-effect distinguishing trastuzimab from lapatinib [38]. The omission of diarrhoea on statistical grounds, in this case, would result in the loss of potentially important information from the utility function.

The specific measurement model tested, determined by a panel of experts belonging to the MAUCa Consortium, was not the only model that could have been posited. Although based on the well-established factor structure of the QLQ-C30, there is arguably a causal structure amongst the domains [28] that could be represented by a structural model (rather than mere correlations amongst the domains, as in the present analysis). To date, the function of factor analysis in the development of health state classification systems has been to establish what domains exist amongst the items; causal structure has not been considered. However, this issue is of practical significance in that one domain may be causally related to another, and if an item is selected from each domain this violates the

principle of structural independence, potentially resulting in double-counting in the utility instrument. A similar problem arises if two or more dimensions are reflective of a higher-order dimension.

Conclusions

A confirmatory approach to determining dimensionality for the construction of a health state classification system was found to be more efficient and to produce a more readily interpretable domain structure for the QLQ-C30. The confirmatory aspect of this prototype analysis will now be applied on a much larger scale as part of the MAUCa project, involving the pooling of a large number of international data sets covering a range of countries, cancer sites and stages. Based on the results, a definitive health state classification system will be determined. This will pave the way for valuation surveys which will provide country-specific utility weights for this health state classification system, and thereby complete the provision of a preference-based measured derived from the QLQ-C30.

Competing interests
None

Authors' contributions
DC performed statistical analysis and wrote the manuscript.
MK contributed to the conception and design of the research and co-wrote the manuscript.
NA contributed to the conception and design of the research.
PF contributed to the conception and design of the research.
PG contributed clinical expertise.
MJ contributed statistical expertise.
JP contributed statistical expertise.
DR assisted in the design of the research.
GV contributed clinical expertise.
RV contributed to the conception and design of the research.
TY assisted in the design of the research.
Members of the MAUCa Consortium contributed to the conception and design of the research and/or assisted in accessing data.

All authors read and approved the final manuscript.

Acknowledgements

The MAUCa Consortium, in addition to those named as authors, consists of the following members, all of whom made some contribution to the research reported in this paper, as outlined above: John Brazier, David Cella, Stein Kaasa, Paul Krabbe, Helen McTaggart-Cowan, Stuart Peacock, Simon Pickard, Neil Scott, Martin Stockler and Deborah Street. This research was supported by a National Health and Medical Research Council (Australia) Project Grant.

References

1. Claxton, K., M. Sculpher, and M. Drummond, *A rational framework for decision making by the National Institute For Clinical Excellence (NICE).* Lancet, 2002. **360**(9334): p. 711-715.
2. Scuffham, P., J. Whitty, and R. Viney, *The use of QALY weights for health state valuation and QALY calculations: A review of industry submissions requesting listing on the Australian Pharmaceutical Benefits Scheme 2002 to 2004.* Pharmacoeconomics, 2008. **26**(4): p. 297-310.
3. Drummond, M.F., et al., *Methods for the economic evaluation of health care programmes*. 3rd ed. Oxford medical publications. 2005, Oxford: Oxford University Press. xv, 379 p.
4. Brazier, J.E., et al., *Estimating a preference-based index from the Japanese SF-36.* Journal of Clinical Epidemiology, 2009. **62**(12): p. 1323-31.
5. Ware, J.E. and C.D. Sherbourne, *The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection.* Medical Care, 1992. **30**(6): p. 473-483.
6. Aaronson, N.K., et al., *The European Organisation for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology.* Journal of the National Cancer Institute, 1993. **85**(5): p. 365-376.
7. Ware, J.E., et al., *SF-36® Health Survey Manual and Interpretation Guide*. 1993, Boston: New England Medical Center, The Health Institute.
8. Cella, D., *Manual of the Functional Assessment of Chronic Illness Therapy (FACIT) scales (version 4).* 1997, Evanston: Northwestern University.
9. de Bekker-Grob, E.W. and M.G. Ryan, K. (2010) *Discrete choice experiments in health economics: a review of the literature*. Health Economics **Volume**, DOI: 10.1002/hec.1697
10. Dolan, P., *Modelling valuation for Euroqol health states.* Medical Care, 1997. **35**: p. 351-363.
11. Norman, R., et al., *International Comparisons in Valuing EQ-5D Health States: A Review And Analysis.* Value in Health, 2009. **12**(8): p. 1194-1200.
12. Feeny, D., et al., *Multi-attribute health status classification systems. Health Utilities Index.* Pharmacoeconomics, 1995. **7**(6): p. 490-502.
13. Feeny, D.H., G.W. Torrance, and W.J. Furlong, *Health Utilities Index*, in *Quality of Life and Pharmacoeconomics in Clinical Trials*, B. Spilker, Editor. 1996, Lippincott-Raven Press: Philadelphia. p. 239-252.
14. Brazier, J.E., J. Roberts, and M. Deverill, *The estimation of a preference-based measure of health from the SF-36.* Journal of Health Economics, 2002. **21**: p. 271-292.
15. Brazier, J.E., et al., *Should patients have a greater role in valuing health states?* Applied Health Economics & Health Policy, 2005. **4**(4): p. 201-8.
16. Brazier, J.E., et al., *Estimation of a preference-based index from a condition-specific measure: the King's Health Questionnaire.* Medical Decision Making, 2008. **28**(1): p. 113-26.
17. Brazier, J.E., 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 Technology Assessment, 2012 forthcoming.
18. Brazier, J.E., et al., *Deriving a preference-based single index from the UK SF-36 Health Survey.* Journal of Clinical Epidemiology, 1998. **51**(11): p. 1115-28.
19. Young, T.A., et al., *The first stage of developing preference-based measures: constructing a health-state classification using Rasch analysis.* Quality of Life Research, 2009. **18**(2): p. 253-265.
20. Feeny, D., *The utility approach to assessing population health*, in *Summary measures of population health: Concepts, ethics, measurement and applications*, J.S. C. Murray, C. Mathers, & A. Lopez, Editor. 2002, World Health Organisation: Geneva. p. 515-528.
21. Young, T.A., et al., *The use of Rasch analysis in reducing a large condition-specific instrument for preference valuation: the case of moving from AQLQ to AQL-5D.* Medical Decision Making, 2011. **31**(1): p. 195-210.

22. Cella, D., et al., *Development and validation of eleven symptom indexes to evaluate response to chemotherapy for advanced cancer.* Journal of the National Comprehensive Cancer Network, 2011. **9**(3): p. 13-24.
23. Cella, D., et al., *What are the most important symptom targets when treating advanced cancer? A survey of providers in the National Comprehensive Cancer Network (NCCN).* Cancer Investigation, 2003. **21**(4): p. 526-535.
24. Young, T., et al., *The Use of Rasch Analysis as a Tool in the Construction of a Preference Based Measure: The Case of AQLQ. Health Economics and Decision Science Discussion Paper 07/01, School of Health and Related Research, The University of Sheffield.* http://www.shef.ac.uk/content/1/c6/01/87/47/HEDS%20DP%200701.pdf. 2007.
25. Rowen, D., et al., *Deriving a preference-based measure for cancer using the EORTC-QLQC30.* Value in Health, 2011. **14**(5): p. 721-731.
26. Bjordal, K., et al., *A 12 country field study of the EORTC QLQ-C30 (version 3.0) and the head and neck cancer specific module (EORTC QLQ-H&N35) in head and neck patients.* European Journal of Cancer. **36**: p. 1796-1807.
27. Kaasa, S., et al., *Prospective randomised multicenter trial on single fraction radiotherapy (8 Gy x1) versus multiple fractions (3 Gy x10) in the treatment of painful bone metastases.* Radiotherapy and Oncology, 2006. **29**: p. 278-284.
28. Gundy, C.M., et al., *Comparing higher order models for the EORTC QLQ-C30.* under review.
29. Tabachnick, B.G. and L.S. Fidell, *Using Multivariate Statistics*. 5th ed. 2007, Boston: Allyn and Bacon.
30. Horn, J.L., *A rationale and test for the number of factors in factor analysis.* Psychometrika, 1965. **30**: p. 179-185.
31. Ostlund, U., et al., *What symptom and functional dimensions can be predictors for global ratings of overall quality of life in lung cancer patients?* Supportive Care in Cancer. **15**: p. 1199-1205.
32. Kline, R.B., *Principles and Practice of Structural Equation Modeling*. 2nd ed. 2005, New York: Guilford Press.
33. McDonald, R.P., *Test Theory: A Unified Treatment*. 1999, New Jersey: Lawrence Erlbaum.
34. Rasch, G., *Probabilistic models for some intelligence and attainment tests*. 1960, Chicago: University of Chicago Press.
35. Pallant, J.F. and A. Tennant, *An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS).* British Journal of Clinical Psychology, 2007. **46**: p. 1-18.
36. Scott, N.W., et al., *Comparing Translations of the EORTC QLQ-C30 Using Differential Item Functioning Analyses.* Quality of Life Research, 2006. **15**(6): p. 1103-115.
37. Scott, N.W., et al., *EORTC QLQ-C30 Reference Values*. 2008: EORTC Quality of Life Group.
38. Tomasello, G., et al., *Jumping higher: is it still possible? The ALTTO trial challenge.* Expert Review of Anticancer Therapy, 2008. **8**(12): p. 1883-1890.

Table 1. The 30 items of the QLQ-C30 and the scales† to which they belong

Item #	Item stem wording	Scale
1	Do you have any trouble doing strenuous activities, like carrying a heavy shopping bag or a suitcase?	PF
2	Do you have any trouble taking a long walk?	PF
3	Do you have any trouble taking a short walk outside of the house?	PF
4	Do you need to stay in bed or a chair during the day?	PF
5	Do you need help with eating, dressing, washing yourself or using the toilet?	PF
6	Were you limited in doing either your work or other daily activities?	RF
7	Were you limited in pursuing your hobbies or other leisure time activities?	RF
8	Were you short of breath?	Dyspnoea (S)
9	Have you had pain?	Pain
10	Did you need to rest?	Fatigue
11	Have you had trouble sleeping?	Insomnia (S)
12	Have you felt weak?	Fatigue
13	Have you lacked appetite?	Appetite loss (S)
14	Have you felt nauseated?	Nausea/vomiting
15	Have you vomited?	Nausea/vomiting
16	Have you been constipated?	Constipation (S)
17	Have you had diarrhea?	Diarrhoea (S)
18	Were you tired?	Fatigue
19	Did pain interfere with your daily activities?	Pain
20	Have you had difficulty in concentrating on things, like reading a newspaper or watching television?	CF
21	Did you feel tense?	EF
22	Did you worry?	EF
23	Did you feel irritable?	EF
24	Did you feel depressed?	EF
25	Have you had difficulty remembering things?	CF
26	Has your physical condition or medical treatment interfered with your family life?	SF
27	Has your physical condition or medical treatment interfered with your social activities?	SF
28	Has your physical condition or medical treatment caused you financial difficulties?	Financial difficulties (S)
29	How would you rate your overall health during the past week?	Global
30	How would you rate your overall quality of life during the past week?	Global

† There are five multi-item functioning scales (PF = physical functioning, RF = role functioning, CF = cognitive functioning, EF = emotional functioning, SF = social functioning); three multi-item symptom scales (Fatigue, Pain, Nausea/vomiting); a global health/quality of life scale; and six single item scales (S).

Figure 1. Frequencies of endorsement of resposnse categories for the27 items of the QLQ-C30 included in the analyses. Item content matching each item number can be found in Table 1.

Table 2a. Summary of item statistics based on the dimensions established using exploratory factor analysis (EFA). Rasch statistics are those obtained from the final analyses, i.e., those with misfitting items removed.

	Factors and loadings (EFA)[a]			Rasch			
Item	Factor 1	Factor 2	Factor 3	Location	Item fit	DIF[b]	LD[c]
1	0.65	0.00	0.01	-0.92	0.88	Gender, site[d]	
2	0.75	-0.07	-0.07	-1.23	-1.34		3
3	0.82	-0.06	-0.15	0.96	0.06		2
4	0.65	0.03	0.14	0.14	0.73		10
5	0.52	-0.02	-0.05	3.40	0.95		
6	0.86	-0.08	0.02	-0.81	-3.89	Site	7
7	0.77	0.02	-0.02	-0.51	-1.37		6
8	0.15	0.17	0.08	Not included in Rasch analysis (weak factor loadings)			
9	0.49	0.06	0.05	Misfit			
10	0.53	0.26	0.25	-0.42	0.59		4
11	-0.02	0.34	0.02	Misfit			
12	0.25	0.38	0.34	-0.97	-0.37		18
13	0.01	0.12	0.62	-0.05	0.34		
14	-0.11	0.03	0.83	0.63	0.28	Site	15
15	-0.11	-0.01	0.77	1.64	-1.73		14
16	0.14	-0.03	0.18	Not included in Rasch analysis (weak factor loadings)			
17	0.02	-0.07	0.29	Not included in Rasch analysis (weak factor loadings)			
18	0.24	0.36	0.36	-1.26	-0.05		12
19	0.68	0.11	-0.02	-0.62	1.51		
20	0.15	0.41	0.29	Misfit			
21	-0.01	0.83	-0.10	0.28	0.29	Gender	
22	-0.01	0.91	-0.23	Misfit		Gender, site	
23	-0.02	0.60	-0.01	-0.28	-1.31		
24	-0.05	0.77	0.04	-0.001	1.49		
25	0.08	0.27	0.27	Not included in Rasch analysis (weak factor loadings)			
26	0.15	0.35	0.09	Misfit			
27	0.42	0.24	0.13	Misfit			

[a] principal axis factoring extraction, direct oblimin rotation

[b] DIF = differential item functioning. Grouping variables exhibiting DIF for the item are listed in this column.

[c] LD = local dependency. Values in this column represent numbers of items with which the item has a residual correlation following Rasch analysis

[d] Cancer sites included prostate, breast, lung and other

Table 2b. Summary of item statistics based on the dimensions established using confirmatory factor analysis (CFA). Rasch statistics are those obtained from the final analyses, i.e., those with misfitting items removed.

	A priori factors, guided by	CFA	Rasch			
Item	conceptual model	Loadings	Location	Item fit	DIF[a]	LD[b]
1	Physical functioning	0.68	-1.01	0.79	Gender, site	
2	Physical functioning	0.73	-1.33	-1.64		3
3	Physical functioning	0.77	0.90	-0.13		2
4	Physical functioning	0.72	0.06	0.71		10
5	Physical functioning	0.49	3.39	0.76		
6	Role functioning	0.90	-0.90	-3.62	Site	7
7	Role functioning	0.85	-0.59	-0.48		6
9	Pain	0.64	-0.38	1.42		
10	Fatigue	0.84	-0.52	0.97		4
12	Fatigue	0.79	0.22	-0.07	Site	
14	Nausea & vomiting	0.93	-1.20	-0.72	Site	
15	Nausea & vomiting	0.75	1.20	-0.21	Site	
18	Fatigue	0.81	-0.22	0.38		
19	Pain	0.94	0.38	0.24		
20	Cognitive functioning	0.82	-0.10	0.52		
21	Emotional functioning	0.82	0.28	0.29	Gender	
22	Emotional functioning	0.86			Gender, site	
23	Emotional functioning	0.59	-0.28	-1.31		
24	Emotional functioning	0.78	-0.001	1.49		
25	Cognitive functioning	0.58	0.10	1.15		
26	Social functioning	0.58	0.323	1.03		
27	Social functioning	0.81	-0.323	0.71	Site	

[a] DIF = differential item functioning. Grouping variables exhibiting DIF for the item are listed in this column.

[b] LD = local dependency. Values in this column represent numbers of items with which the item has a residual correlation following Rasch analysis

[d] Cancer sites included prostate, breast, lung and other

Table 3. Correlations between factors obtained from the confirmatory factor analysis.

	PF	RF	EF	SF	CF	Pain	Fatigue
RF	.855						
EF	.304	.286					
SF	.617	.588	.532				
CF	.437	.437	.56	.627			
Pain	.683	.738	.365	.562	.394		
Fatigue	.698	.646	.564	.656	.752	.572	
NV	.189	.239	.198	.333	.439	.263	.489