
CHERE
CENTRE FOR HEALTH ECONOMICS
RESEARCH AND EVALUATION


UNIVERSITY OF
TECHNOLOGY SYDNEY

UTS:CHERE

The Centre for Health Economics Research and Evaluation (CHERE) was established in 1991. CHERE is a centre of excellence in health economics and health services research. It is a joint Centre of the Faculties of Business and Nursing, Midwifery and Health at the University of Technology, Sydney, in collaboration with Central Sydney Area Health Service. It was established as a UTS Centre in February, 2002. The Centre aims to contribute to the development and application of health economics and health services research through research, teaching and policy support. CHERE's research program encompasses both the theory and application of health economics. The main theoretical research theme pursues valuing benefits, including understanding what individuals value from health and health care, how such values should be measured, and exploring the social values attached to these benefits. The applied research focuses on economic and the appraisal of new programs or new ways of delivering and/or funding services. CHERE's teaching includes introducing clinicians, health services managers, public health professionals and others to health economic principles. Training programs aim to develop practical skills in health economics and health services research. Policy support is provided at all levels of the health care system by undertaking commissioned projects, through the provision of formal and informal advice as well as participation in working parties and committees.

University of Technology, Sydney
City campus, Haymarket
PO Box 123 Broadway NSW 2007
Tel: +61 2 9514 4720
Fax: + 61 2 9514 4730
Email: mail@chere.uts.edu.au
www.chere.uts.edu.au

Validity, reliability and responsiveness of the EORTC QLQ-C30 and the EORTC QLQ-LC13 in Australians with early stage non-small cell lung cancer

Madeleine King[1], Julie Winstanley[2], Patsy Kenny[1], Rosalie Viney[1], Siggi Zapart[1], Michael Boyer[3]

CHERE WORKING PAPER 2007/13

1. Centre for Health Economics Research and Evaluation (CHERE)
Faculty of Business
University of Technology, Sydney

2. Osman Consulting, Sydney

3. Royal Prince Alfred Hospital, Sydney

First Version: December 2007
Current Version: December 2007

Abstract

Aim: To assess the validity, reliability and responsiveness of two questionnaires, the QLQ-C30 and LC-13, as measures of health-related quality of life (HRQOL) in an Australian sample of people with early stage non-small cell lung cancer.

Background: These two questionnaires are complementary components of the European Organisation for Research and Treatment of Cancer's (EORTC's) modular approach to measuring HRQOL: the QLQ-C30 is the core questionnaire, containing 30 items relevant to all cancers; the QLQ-LC13 contains 13 items specific to lung cancer.

Methods: These two complementary questionnaires were assessed with data obtained from 183 participants of a randomised control trial investigating the use of Positron Emission Tomography in the management of stage I or II non-small cell lung cancer. A cohort of 173 participants, were treated by surgery and then followed for two years. Participants completed HRQOL questionnaires before the PET scan, before and after surgery, one month after surgery, and then four monthly for two years. Construct validity was tested with confirmatory factor analysis and correlation analysis was used to test for convergent/divergent validity. Discriminant validity was tested by assessing the sensitivity of the scales to the effects of moving from early to late stage disease, asymptomatic to mildly symptomatic, and to the effects of age, gender and number of comorbitities. Mean differences (standardized response means (SRM)) and effect sizes were estimated for: patients with Stage 1/11 and metastatic disease; ECOG score 0 and ECOG score 1; older and younger patients; men and women; patients with no comorbidities and those with 1 or more comorbidities. Reliability was assessed in terms of internal consistency and test-retest reliability. Responsiveness to the effects of major thoracic surgery, adjuvant radiotherapy, and disease recurrence was assessed by estimating mean differences (standardized response meansSRM's and effect sizes for patients who underwent surgery, radiotherapy and whose disease recurred, respectively.

Results: The factor structure reported previously was replicated in this sample, confirming the questionnaires' construct validity. Most scales demonstrated good to excellent internal consistency (Cronbach's alpha range: 0.86 – 0.94); the exceptions were the cognitive function (0.68) and nausea/vomiting scales (0.67). Test-retest reliability was generally good (intraclass correlation (ICC) range: 0.70 – 0.81); the exceptions were the pain and nausea/vomiting scales (ICC 0.56 and 0.42). Most scales were sensitive to the large effect of moving from early to later stage disease with (SRM range: 21.3 – 54.0; effect size range:1.14 – 1.97 (except for emotional functioning: 13.7; 0.60)). The scales were also sensitive to small effects, detecting small to moderate differences for age (large for social functioning) and comorbidities, and small differences for moving from asymptomatic to mildly symptomatic disease, and for age. Responsiveness was also confirmed with most scales responsive to the large expected effects of surgery and disease progression (SRM range: 21.6 – 41.4; effect size range: 0.94 – 1.89 (emotional functioning: 5.5; 0.19)).

Conclusions: The QLQ-C30 and QLQ-LC13, when used together, provide a valid, reliable and responsive measure of HRQOL in Australians with early stage non-small cell lung cancer.

Acknowledgments
This research was supported by an Australian National Health and Medical Research Council Project Grant. We would like to thank Christine Pollicino, Jocelyn McLean, Michael Fulham and Brian McCaughan for assistance with patient recruitment and data collection, and the patients themselves for their important contribution in completing questionnaires in the months and years after their surgery.

Disclosures
This paper is original research and has not been published elsewhere. Two papers published in the Journal of Clinical Oncology (Viney et al 2004; 22: 2357-62; Kenny et al 2008 26(2): 233-241) presents a different components of the same research project.

The research has been presented in part at two conferences: the Australian Health Outcomes Conference (Canberra, November 2005) and the annual conference of the Clinical Oncology Society of Australia (Brisbane, November 2005). The abstract for latter presentation was published in the Asia-Pacific Journal of Clinical Oncology 2005; 1(Suppl): A48.

The authors do not have any conflicts of interest to declare.

INTRODUCTION and BACKGROUND

Health-related quality of life

In its broadest sense, the term *quality of life* covers aspects of life that are beyond the scope of health care, such as living standards, housing, education, employment and the environment. In the context of health, its meaning is restricted to aspects of welfare that relate to health and health care (Ware 1987; Schipper, Clinch et al. 1996). The term health-related quality of life (HRQOL) is often used to differentiate the restricted from the broader sense (Wood-Dauphinee 1999). There is no single, concise definition of HRQOL, but it is widely accepted in the context of health research that HRQOL reflects the impact of disease and treatment on a person's ability to function physically, socially and emotionally, and their symptom experience.

Measurement of HRQOL

The formal assessment of the health and well-being of individuals and populations is referred to as health status assessment or HRQOL assessment. There are many instruments that measure the impact of disease and treatment on the HRQOL of patients. The appropriate instrument for a particular application is determined by the purpose of the measurement and the kind of information required (Osoba, Aaronson et al. 1991; Guyatt, Jaeschke et al. 1996). Some are specific to a particular disease or treatment, others are general.

All measurement instruments, regardless of whether they consist of a single item, such as the response to a single global question, or have multiple items with multi-item scales and summary scores, should satisfy basic properties if they are to be considered useful. These properties or attributes include validity, reliability, sensitivity and responsiveness. They are interrelated, yet each is independently important.

Validity

The validity of an instrument is "the degree to which the instrument measures what it purports to measure" (Lohr, Aaronson et al. 1996). Although validity is often considered an attribute of an instrument, its true significance relates to the uses to which the instrument is put (Nunnally 1978; Fayers and Machin 2000) and the inferences that are drawn from resultant scores (Streiner and Norman 1996). An instrument should be validated for every intended purpose.

Validity can be subdivided into three main aspects, namely content, criterion and construct validity. However, the methodology and terminology of validity are complex and there is often overlap in the meanings of terms and the evidence provided by various methods.

Content validity is the extent to which the items of the instrument cover the range of issues that are relevant to its intended use (Fayers and Machin 2000). The more representative the sample of items, the more likely the instrument will yield inferences that hold true in a wide range of circumstances (Streiner and Norman 1996). The wording of items and response scales should be unambiguous, and redundancy should be minimised. *Face validity* is a closely related concept. The distinction is that content validity is determined during the development of an instrument while face validity is a used as criterion when choosing among existing instruments for a specific purpose (Fayers and Machin 2000).

Construct validity is the extent to which the relationships observed among variables conform with hypothesised relationships (Streiner and Norman 1996; Fayers and Machin 2000). There are two main types of evidence. The first tests hypothesised relationships among latent variables. Evidence is generally sought by correlation of observed variables: correlations between items in the same scale, correlations between an item and items in other scales, correlations between a scale score and its constituent items, and correlations among scales of one or more instruments. *Convergent validity* is supported by correlation among measures of latent variables that are hypothesised to be similar. *Discriminant validity* is supported by lack of correlation among measures of latent variables that are hypothesised to be dissimilar. Common methods of analysis include *factor analysis, path analysis, multitrait-scaling analysis and multitrait-multimethod analysis*.

The second form of evidence supports hypothesised relationships between latent variables and external criteria. For example, patients with early stage cancer may be expected to have better QOL than patients with advance cancer. This type of evidence is said to support *clinical validity* or "*known-groups*" *validity* because groups of patients are often defined by clinical criteria. This also provides evidence of the *sensitivity* of a scale to clinically important differences. This has also been called *discriminative validity* (Stockler, Osoba et al. 1999).

Criterion validity is the extent to which a measure corresponds to an external criterion (Nunnally 1978; Streiner and Norman 1996; Fayers and Machin 2000). Criterion validity can be divided into two aspects. *Concurrent validity* means agreement with a true value, or "gold standard", which does not exist for QOL. If a short version of an established questionnaire is being developed, the long version may be considered the standard. *Predictive validity* concerns the ability of an instrument to predict future health status or future events (such as hospitalisation or death).

Reliability

The reliability of an instrument is its ability to yield reproducible and consistent results (Fayers and Machin 2000). Formal definitions of reliability involve notions of random variation or measurement error. In QOL assessment, random variations may include real but transient variations in health or circumstance or in the perception of health or circumstance. Measurement error may be due to scale coarseness in approximating the continuous latent variable and inconsistent use of the scale by the respondent.

The consistency of the items in a multi-item scale as measures of the latent variable is called *internal consistency*. There are many measures of internal consistency, *Cronbach's alpha coefficient* being the most commonly used (Hattie 1985). Internal consistency is in many senses a form of validity (Fayers and Machin 2000) and is commonly estimated and reported along with construct validity.

Another type of reliability is the *reproducability* or stability of scores on a scale when the circumstances of assessment differ but the patient's QOL does not. Circumstances may include mode of administration, place of completion, and time (repeated occasions).

Reproducability across repeated measurements is commonly called *test-retest reliability*. One of the difficulties with estimating test-retest reliability for QOL measures is identifying the appropriate patient population and period. The period must be long enough so respondents do not remember their responses to questions, and short enough so that their QOL has not changed. Reproducability is often assessed with *kappa* for discrete data, or the *intra-class correlation coefficient (ICC)* for continuous data (Dunn 1992).

Responsiveness

The *responsiveness* of an instrument is its ability to detect change (Lohr, Aaronson et al. 1996). Responsiveness is tested and calibrated in situations where clinically meaningful change is likely to occur, such as a treatment of known effectiveness administered to patients who are likely to respond (Liang, Larson et al. 1985; Deyo, Diehr et al. 1991) or disease progression as part of the natural history of a disease. The most commonly used responsiveness measures are the mean change (standardised response mean) proposed by Liang et al (1990) and effect size, proposed by Kazis et al (1989).

Measuring HRQOL in cancer care

Traditionally, the evaluation of cancer treatments focused on biomedical outcomes such as tumor response, time to progression, disease free and overall survival rate and treatment related toxicities. However, it has now become increasingly accepted that the measurement of HRQOL needs to be included if we are to obtain a comprehensive evaluation (Osoba 1995).

The measurement of HRQOL in the area of cancer care is particularly important because the physical and psychological effects of the disease, and the benefits and toxicities of cancer treatments (Schipper and Clinch 1988; Fallowfield 1990; Cella and Tulsky 1993b; Cleton 1995; Greco 1995; Hanks and Hoskin 1995; Maguire 1995; Steel 1995) have a direct effect on patients' wellbeing and their ability to perform their usual roles and abilities. There are several instruments that measure HRQOL in this area. One of the most recommended instruments is the European Organization for Research and treatment of Cancer Quality of Life Questionnaire Core module (EORTC QLQ-C30) (Aaronson, Ahmedzai et al. 1993).

The QLQ-C30 was developed by the EORTC Study Group for Quality of Life (Aaronson, Bullinger et al. 1988; Aaronson, Ahmedzai et al. 1991a; Aaronson, Ahmedzai et al. 1993). It is the core component of the EORTC's modular approach to QOL assessment and represents QOL domains relevant across a wide range of cancer sites and treatment types. The QLQ-C30 is complemented by modules specific to particular cancers such as lung cancer and breast cancer. The core module facilitates comparison across the diversity of trials administered by the EORTC, and the disease-specific modules provide sensitivity for particular trials (Aaronson, Cull et al. 1996).

From it's beginnings as the first generation QLQ-C36 (Aaronson, Ahmedzai et al. 1991a) to the current version QLQ-C30 v3 (Aaronson, Ahmedzai et al. 1993), it has undergone a continual process of development and validation. It has been found to have good to excellent validity, reliability, sensitivity and interpretability, and has been used in a wide range of cancer trials as well as other various non-trial studies world wide. A bibliography of validation studies can be

found in the EORTC QLQ-C30 scoring manual (Fayers, Aaronson et al. 2001) and a summary of findings from numerous studies can be found in Spilker et al (Spilker 1996).

The lung cancer module, the QLC-LC13, is meant for use with a wide variety of lung cancer patients in varying disease stage and treatment modality (Bergman, Aaronson et al. 1994). This was the first module developed to address specific symptoms associated with a particular cancer and its treatment. It was constructed in parallel with, and validated in field tests together with the QLQ-C30 (Bergman, Aaronson et al. 1994), and has subsequently been validated in other studies (Chie, Yang et al. 2004; Nowak, Stockler et al. 2004).

Ongoing validation of the EORTC QLQ-C30 and the EORTC QLQ-LC13

The validity of a HRQOL instrument is not something that is established by a single or even a few studies. Whether or not the instrument produces sensible and useful results in various circumstances should be judged in an ongoing process of validation (Streiner and Norman 1996; Fayers and Machin 2000). Although the EORTC QLQ-C30 and the EORTC QLQ-LC13 have undergone a continual process of validation across a range of health care contexts and disease groups, and in different nationalities and cultures, our confidence in, and understanding of, the instruments will develop as the body of evidence accrues.

In most instances, HRQOL instruments are validated in studies or trials designed specifically for validation purposes. However, researchers can also use data from existing studies, designed for other purposes, to validate an instrument. An Australian randomized clinical trial (RCT) designed to investigate the role of Positron Emission Tomography (PET) in the management of early stage non-small cell lung cancer (NSCLC) provided a good opportunity to validate the QLQ-C30 and the QLQ-LC13 in this population.

RCT to investigate PET in the management of early stage Non-Small Cell Lung cancer (NSCLC)

NSCLC represents approximately 80% of all lung cancer. Approximately 25% of patients present with what appears to be resectable disease, but relapse is common with up to 40% of patients with Stage 1 and 60% of patients with Stage 2 disease at surgery experiencing disease recurrence following surgical resection. These patients went through the trauma of major thoracic surgery without the cure that was hoped for; in that sense, their thoracotomies were futile.

Positron emission tomography (PET) is a relatively new imaging technology with the potential to improve pre-operative staging. Many malignant tumours show increased glucose utilisation when compared to normal tissues (Nolop, Rhodes et al. 1987). Whole body PET with ^{18}F-fluorodeoxyglucose (FDG) can identify regions of increased glucose metabolism in non-enlarged structures, allowing detection of tumour metastases earlier than with anatomic imaging methods. Data suggest PET may improve the accuracy of pre-operative staging of NSCLC, but, in general, these are from small, retrospective, uncontrolled series (Wahl, Quint et al. 1994; Weder, Schmid et al. 1998; Saunders, Dussek et al. 1999). A recent prospective uncontrolled study reported sensitivity and specificity of PET for detection of mediastinal and distant metastatic disease of 95 per cent and 83 per cent, respectively (Pieterman, van Putten et al. 2000). PET is costly, and resource implications of its widespread use in staging NSCLC are significant. There is

increasing pressure for PET to be included in the standard diagnostic work-up prior to decisions about surgical management of NSCLC (Robert and Milne 1999; Berlangieri and Scott 2000).

In the first randomised controlled trial of patients with a clinical diagnosis of Stage I-II of NSCLC, 184 patients were recruited and randomised. Following exclusion of one ineligible patient, 92 patients were assigned to no-PET and 91 to PET. Of these, 173 were treated by surgery. Compared with conventional staging PET upstaged 22 patients, confirmed staging in 61 and staged two patients as benign. Stage IV disease was rarely detected (2 patients). PET led to further investigation or a change in clinical management in 13% of cases, and provided information that could potentially have affected management in a further 13% of cases. There was no significant difference between the trial arms in the number of thoracotomies avoided ($p=0.2$). It was concluded that for patients who are carefully and appropriately staged as having stage I-II disease, PET provides potential for more appropriate stage-specific therapy, but may not lead to a significant reduction in the number of thoracotomies avoided. A full report of the RCT is available in a paper by Rosalie Viney *et al* (Viney, Boyer et al. 2004).

Further research conducted using the RCT sample

HRQOL data was collected at recruitment from all 183 patients eligible to participate in the RCT. A cohort of 173 participants who were treated by surgery was then followed for two years. During this period, clinical outcomes, resource utilization and HRQOL were measured. The HRQOL data was evaluated from two angles.

The first aim was to describe the HRQOL in this patient group; these analyses are published elsewhere (Kenny, King et al. 2008).

The second aim was to describe the measurement properties of the HRQOL instruments. Specifically, to assess the validity, reliability and responsiveness of the QLQ-C30 and QLQ-LC13, as measures of HRQOL in an Australian sample of people with early stage NSCLC. These analyses are the subject of this discussion paper.

METHODS

Recruitment and data collection

Participants in this study were 183 patients recruited from the practices of six thoracic surgeons in Sydney, Australia between April 1999 and December 2000. The patients were participating in a randomised controlled trial to investigate the impact of Positron Emission Tomography (PET) on the clinical management and surgical outcomes for patients with a clinical diagnosis of Stage I or II NSCLC. The trial found no significant difference in management between the intervention and control groups (Viney, Boyer et al. 2004). Of the 183 participants, 173 were treated by surgery and then followed for two years. The study was approved by the relevant institutional ethics committees.

The EORTC QLQ-C30 version 3 (Aaronson, Ahmedzai et al. 1993), and the EORTC QLQ-LC13 (Bergman, Aaronson et al. 1994) were used to measure HRQOL. Questionnaires were self-completed at recruitment to the study, at hospital admission and discharge, one month and four months after surgery and then every four months until two years after surgery. Additional

assessments were completed at the beginning and end of adjuvant therapy and participants diagnosed with recurrent disease were asked to complete monthly assessments for as long as possible, in order to capture changes within the four-month period. Preoperative and discharge assessments were completed at the clinic or hospital and the remaining assessments were completed by postal survey. Socio-demographic characteristics were collected at the recruitment interview and clinical information was collected from individual hospital medical records, from the surgeon and from the patient's general practitioner.

Questionnaires

The EORTC QLQ-C30 and the EORTC QLQ-LC13 were chosen because they were developed in a sound, rigorous way, and had been shown to be valid and reliable. The QLQ-C30 core questionnaire is a generic cancer instrument that contains 30 items relevant to all cancers. The QLQ-LC13 is a lung cancer module that contains 13 items specific to lung cancer.

The QLQ-C30 incorporates nine multi-item scales: five functional scales (Physical – five items, Role – two items, Emotional – four items, Social – two items, Cognitive functioning – two items): three symptom scales (Fatigue – three items, Pain -two items, and Nausea and Vomiting – two items); and a Global Health Status/QOL scale- two items. It also includes six single items that assess symptoms commonly reported by cancer patients (Dyspnoea, Insomnia, Appetite loss, Constipation, Diarrhea and Financial difficulties). The first 28 items have a four-point Likert response scale, namely 'not at all', 'a little', 'quite a bit' and 'very much'. The last two items, rating overall health and overall quality of life during the past week, have a response scale ranging from 1 (very poor) to 7 (excellent). (Fayers, Aaronson et al. 1999) The questionnaire takes approximately 11-12 minutes to complete. High scores on the functioning scales indicates good functioning, high scores on the symptom scales indicates worse symptoms.

The QLQ-LC13 comprises: one multi item scale, Dyspnoea, which has three items, and nine single items (Coughing, Haemoptysis, Sore mouth, Dysphagia, Peripheral neuropathy, Alopecia, Pain in chest, Pain in arm/shoulder and Pain in other parts). The first 12 items have a four-point Likert response scale, namely 'not at all', 'a little', 'quite a bit' and 'very much'. Item 12 also includes an open ended question following the initial response. Item 13 is also composed of two parts. It begins with a 'yes', 'no' response format, followed by a four-point response scale of 'not at all', 'a little', 'quite a bit' and 'very much' for those who answered 'yes'. (Fayers, Aaronson et al. 1999). As for the QLQ-C30, high scores on functioning scales indicate good functioning, high scores on the symptom scales indicate worse symptoms.

Analysis

The sample baseline socio-demographic and clinical characteristics were described and the proportions receiving subsequent therapies, diagnosed with lung cancer recurrence and dying during follow-up were reported. A range of analyses was carried out to confirm the validity, reliability and responsiveness of the EORTC QLQ-C30 and the EORTC QLQ-LC13 questionnaires. The main focus of these analyses is on the multi-item scales, although the single items scores are used for some analyses.

The QLQ-C30 and QLQ- LC13 items were summarised into scales as per the scoring manual (Fayers, Aaronson et al. 2001). All multi-item scales are the mean score of the relevant items

transformed to a score between 0 and 100. A higher score represents better quality of life for the global health status and functional scales and worse quality of life (more symptoms) for the symptom scales.

The pattern and extent of missing assessments were reported for nine time-points (preoperative, hospital discharge, 1, 4, 8, 12, 16, 20 and 24 months after surgery). For those who had surgery, missing assessments were separated by disease status, i.e. recurrence and no recurrence.

Validity

There are several types of validity. A range of analyses was conducted to test the construct, convergent, divergent and discriminate validity of the EORTC QLQ-C30 and the EORTC QLQ-LC13.

Construct validity

Construct validity is reflected in the relationships of the items and the domain scales. These relationships have been specified for both of these instruments by the EORTC, and are summarized as measurement models. Confirmatory factor analysis (CFA), which is a correlation analysis conducted on item-level data, was conducted to test the EORTC measurement models for our sample. The domain scores at the last time-point for each person were used for this analysis.

Various fit indices were produced including the Goodness of Fit Index (GFI)(Joreskog 1993; Tanaka 1993), the Adjusted Goodness of Fit Index (AGFI), which takes into account the degrees of freedom available for testing the model, and the Comparative Fit Index (CFI) (Bentler 1990). In all of these, a value of 1 indicates perfect fit and values over 0.95 for GFI and 0.90 for AGFI and CFI are generally thought to indicate adequate fit (Arbuckle and Wothke 1999; Hu and Bentler 1999). The Chi-square distribution and Root Mean Square Error of Approximation (RMSEA)(Browne and Cudeck 1993), were also calculated. These indices measure how *badly* the proposed model fits the data. For the RMSEA, a value of 0.00 indicates a perfect fit, values of 0.08 or less indicate a reasonable fit and 0.05 or less a well fitting model (Browne and Cudeck 1993). The internal consistency of each domain scale was measured using the Cronbach's alpha statistic. Values above 0.7 are generally regarded as acceptable, over 0.8 good, and over 0.9 excellent (Fayers and Machin 2000).

Convergent and Divergent validity

Convergent and divergent validity is tested by *a priori* expectations about the relationships among domain scales. In this sample it was expected that there would be a moderate correlation among the physical-based functioning scales and the symptom scales, and among the psychosocial scales. The correlation among the physical and psychosocial scales was expected to be lower. Correlation analysis was conducted on domain-level and single item data (from the last timepoint for each person), specifically the correlation matrix of all multi-item scales and single items of the QLQ-C30 and the QLQ- LC13. Correlations from 0.10 to 0.29 are generally regarded as small, 0.30 to 0.49 as medium, and 0.50 or more as large (or strong) (Cohen 1988).

Discriminant validity

Discriminant validity is tested by *a priori* expectations about groups known to differ in clinically relevant ways. The following *a priori* expectations were specified for this sample:

1) patients with newly diagnosed Stage I or II disease would have much better HRQOL across all domains than patients with metastatic disease;

2) patients who were asymptomatic at recruitment (as recorded by a clinician-rated ECOG performance status of zero) would have slightly better HRQOL across all domains than patients who were symptomatic at recruitment but whose symptoms had little or no impact on their daily function (as recorded by a clinician-rated ECOG performance status of one);

3) at recruitment, the HRQOL of older patients would be lower than for younger patients for all physical but not psychosocial domains;

4) at recruitment, men would have a slighter better HRQOL (about 5 points) than women;

5) HRQOL at recruitment would be inversely related to the number of comorbidities.

The first expectation constitutes a clinically "large" effect while the other expectations constitute a clinically "small" effect. Mean differences (in HRQOL scale units) and effect sizes (mean difference divided by the between-person standard deviation) were estimated to test these expectations.

The sensitivity of the scales to the large effects of moving from early to late stage disease was tested with the HRQOL data at recruitment for the 113 patients whose disease did not progress contrasted with the HRQOL data from the last observation for the 45 patients whose disease did progress. The sensitivity of the scales to the small effect of moving from asymptomatic to mildly symptomatic was tested with the HRQOL data at recruitment of patients with ECOG 0 status contrasted with patients with ECOG 1 status.

HRQOL data at recruitment was also used to test the sensitivity of the scales to gender, age and comorbidities. For all expectations, significance testing was also conducted to determine if the differences were significantly different. All analyses were conducted on domain level data. Confidence intervals were determined for all estimates. King has suggested the following guidelines for determining the significance of the mean differences when the QLQ-C30 and QLQ-LC13 are used: For all but three scales (role, emotional and cognitive functioning), a difference of up to 2 points is unlikely to have clinical relevance ("trivial"), a difference of about 5 points is relatively small but may be clinically important ("small"), a difference of about 10 points is likely to have clinical significance ("moderate"), and a difference of 15 or more is relatively large and has clear clinical relevance ("large") (King 1996). For the role functioning scale, the values are 5, 10, 15 and 25. There is insufficient evidence to judge the relative effect size for the emotional and cognitive functioning scales. The empirical effect sizes are generally similar to Cohen's guidelines for small (0.2), medium (0.5) and large (0.8) (Cohen 1988).

Reliability

Test-retest reliability, which assesses measurement stability over time, is relevant to many health research applications because we often need to detect clinically important change over time and we need to be sure that the degree of change detected is greater than that expected by chance. Despite its importance, there is relatively little empirical evidence about the test-retest reliability of HRQOL instruments generally, and these two instruments in particular. Assessing test-retest reliability requires a sub-sample of patients and time points where HRQOL is expected to remain stable. Therefore HRQOL data at recruitment to the RCT and at admission to hospital, which was generally about a week later, was used. Data from participants who had a positive PET scan or any other clinically relevant episode between recruitment and admission, were excluded from the analysis, as any such events may have altered some aspects of their HRQOL.

The summary measure for test-retest reliability is the intra-class correlation coefficient, which is derived from a repeated measures ANOVA. The mean change in the test-re-test data was also examined, as this suggests the degree of change that may be expected by chance for each scale. *Internal consistency*, which assesses the degree of correlation of the items with a multi-item scale, was measured with cronbach's alpha. All analyses were conducted on domain level data. Confidence intervals were determined for all estimates. ICC's of 0.70 were considered acceptable (Fayers and Machin 2000).

Responsiveness

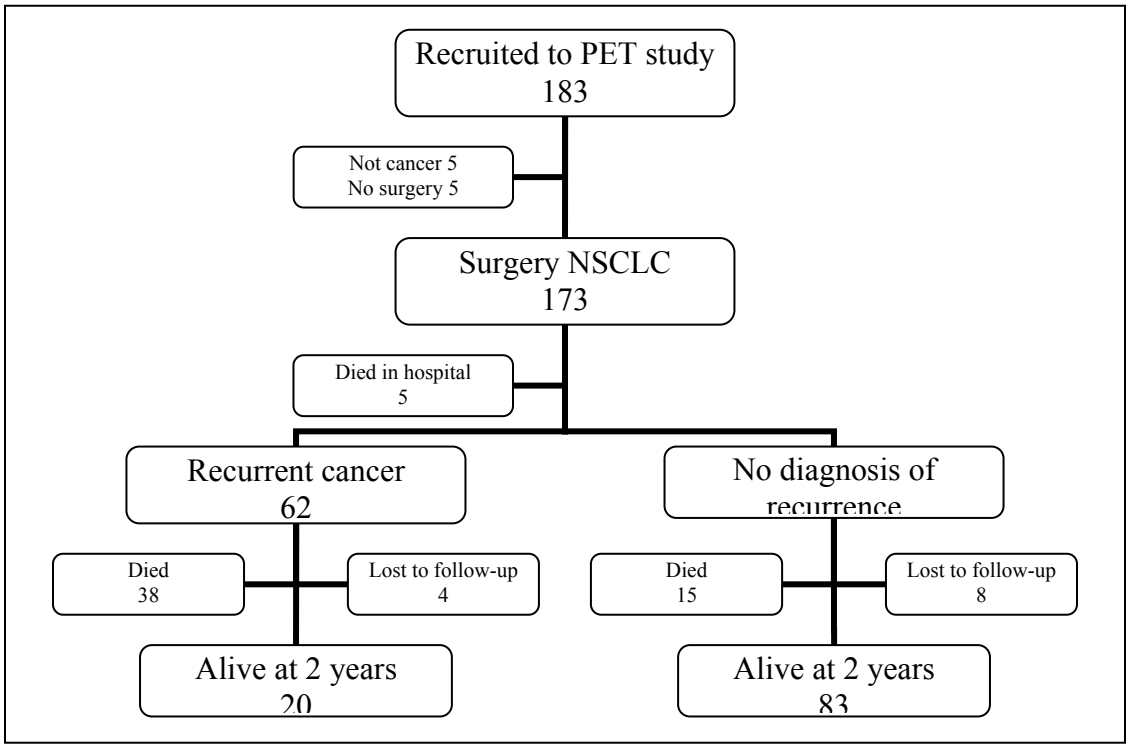
Responsiveness is the ability of an instrument to detect clinically important change. It is a key measurement property for any instrument that is used to evaluate the effect of health interventions on HRQOL in longitudinal studies. A scale that fails to detect a clinically important change when it occurs is worse than useless. Despite its importance, there is relatively little empirical evidence about the responsiveness of HRQOL instruments generally, and these two instruments in particular. Instrument responsiveness can be assessed and characterized by identifying groups of patients whose clinical status is known to have changed during a given time period and considering the change in their HRQOL scores over that period. This sample provided the opportunity to assess responsiveness to the well-known clinical effects of surgery, disease recurrence and adjuvant radiotherapy. The relevant analyses are similar to those for discriminant validity, but in this case the parameter of interest is the mean change over time and the effect size is calculated as the mean change divided by the between-person standard deviation (at the first of the two observations). This version of effect size is sometimes called the standardized response mean.

HRQOL data at the following time points were used to estimate the mean change and effect sizes: surgery - admission and discharge; disease recurrence - recruitment and first and last observation following recurrence; adjuvant radiotherapy – recruitment and first and last radiotherapy treatment. All analyses were conducted on domain level data. Confidence intervals were determined for all estimates

RESULTS

The sample comprised 183 patients participating in the randomised controlled trial investigating the impact of PET on the clinical management and surgical outcomes for patients with Stage I or II NSCLC. Of these, 173 underwent surgery and had a post-operative diagnosis of lung cancer. Five patients died before discharge, reflecting the small but ultimate risk that surgery poses for some patients. A further 62 patients (37%) experienced disease recurrence within two years and only 20 were still alive two years after surgery. Of the remaining 106 patients, 15 died of other causes, eight were lost to follow-up and 83 were disease free at two years.

Figure 1 Study recruitment, treatment and follow-up



Sociodemographic and clinical characteristics

The sociodemographic, clinical and treatment characteristics of the sample are presented in Table 1. The sociodemographic profile was as expected given the type of cancer. The sample comprised mainly older men who had smoked most of their lives. Education levels were low, few had private health insurance, some had Department of Veterans Affairs cards, most were married and born in Australia.

The clinical profile was also typical. Although all had initially been diagnosed as Stage I or II preoperatively on the basis of clinical signs, almost a quarter were found to have more extensive disease at definitive diagnosis post-surgery. Over a third (36%) experienced disease progression during the two year follow-up period, and about the same number (34%) died. A quarter received adjuvant radiotherapy, a fifth received palliative radiotherapy, and almost a tenth received adjuvant chemotherapy

Table1 Socio-demographic, clinical and treatment characteristics of the sample

Characteristic	Percent (n=183)	Mean ± sd
Age (years)		66 ± 9
Male	73	
Australian born	67	
Speak English at home	87	
Did not complete secondary education	65	
Married/defacto	69	
Private health insurance/DVA[1]	26	
Ever smoked	97	
Years smoked		37 ± 15
Pack-years smocked		52 ± 40
ECOG at baseline:		
0	33	
1	60	
2	6	
3	1	
Weight loss at baseline	25	
Clinical stage before surgery:		
I	92	
II	8	
Clinical stage after surgery:		
IA	13	
IB	38	
IIA	2	
IIB	23	
IIIA	18	
IIIB	5	
IV	1	
Resection complete	96	
Surgery pneumonectomy	25	
Adjuvant radiotherapy	24	
Adjuvant chemotherapy	1	
Palliative radiotherapy	19	
Palliative chemotherapy	9	
Recurrent/advanced disease within 2 years	36	
Died within 2 years:		
All causes	33	
Lung cancer	25	

1 Department of Veteran Affairs (DVA) cover for health care.

Missing data

The number providing data at each time-point declined throughout the follow-up period, due largely to death rather than missing data (Table 2). Ten participants (6%) did not contribute data to the HRQOL analysis (3 had no data and 7 had no postoperative data). Forty-three percent completed all HRQOL assessments and a further 13% completed all assessments until death. Not surprisingly, the attrition rate over the two years of the study was greater for the recurrence group. Of 62 participants diagnosed with recurrence, 94% completed HRQOL preoperatively, 50% at one year and 27% at two years. Of 106 participants without recurrence, 97% completed HRQOL preoperatively, 76% at one year and 74% at two years. The rate of missing data ranged from 6% to 21% for those with recurrence and 3% to 16% for those without disease recurrence (see Table 2).

Distribution of HRQOL scores

Inspection of skewness and kurtosis values, histograms, and box and normality plots, revealed several variables were not normally distributed. Significance testing with data that was not normally distributed was conducted using non parametric methods, namely Kruskal Wallis ANOVA and Mann Whitney U tests.

Table 2 Health related quality of life (HRQOL) responses from recruitment to 24 months post surgery

Assessment #	N (%)	1	2	3	4	5	6	7	8	9
Time-point		Preoperative	Discharge	1 month	4 months	8 months	12 months	16 months	20 months	24 months
No cancer*	5 (3)									
No surgery*	5 (3)									
Postop death	5 (3)									
Recurrence	62 (33)									
Not alive (%)		0	0	1 (2)	3 (5)	11 (18)	23 (37)	28 (45)	33 (53)	38 (61)
HRQOL (%)		58 (94)	55 (89)	54 (87)	46 (74)	40 (65)	31 (50)	29 (47)	24 (39)	17 (27)
Missing (%)		4 (6)	7 (11)	7 (11)	13 (21)	11 (18)	8 (13)	5 (8)	5 (8)	7 (11)
No recurrence	106 (58)									
Not alive (%)		0	0	2 (2)	3 (3)	5 (5)	8 (8)	10 (9)	14 (13)	15 (14)
HRQOL (%)		103 (97)	95 (90)	91 (86)	93 (88)	88 (83)	81 (76)	84 (79)	82 (77)	78 (74)
Missing(%)		3 (3)	11 (10)	13 (12)	10 (9)	13 (12)	17 (16)	12 (11)	10 (9)	13 (12)
Total	183 (100)									

* completed HRQOL assessment at study recruitment

Validity

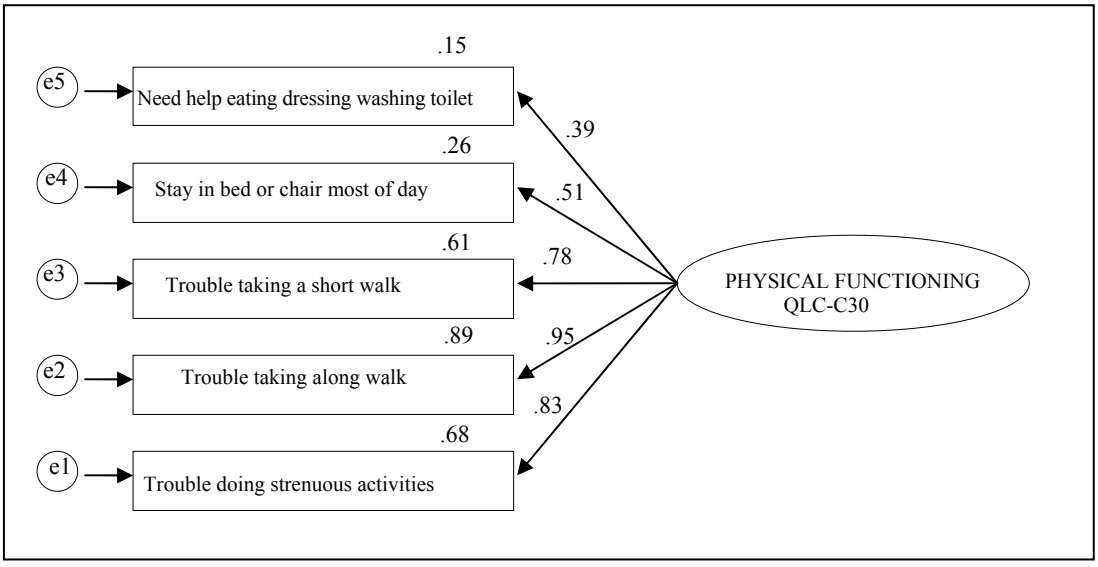
Construct validity

Confirmatory factor analysis (CFA) was conducted on the physical functioning and emotional functioning scales of the EORTC QLQ-C30. CFA was not conducted on the two and three item scales of the EORTC QLQ-C30 and EORTC QLQLC-13 because three has been suggested as the minimum number of items with which to conduct a CFA (Hatcher 1994).

The CFA conducted on the physical functioning and emotional functioning scales of the QLQ-C30 suggest that the measurement models specified by the EORTC for these scales were generally replicated in this sample. Indices of goodness of fit varied but were mostly adequate or better. Internal consistency was shown to be robust.

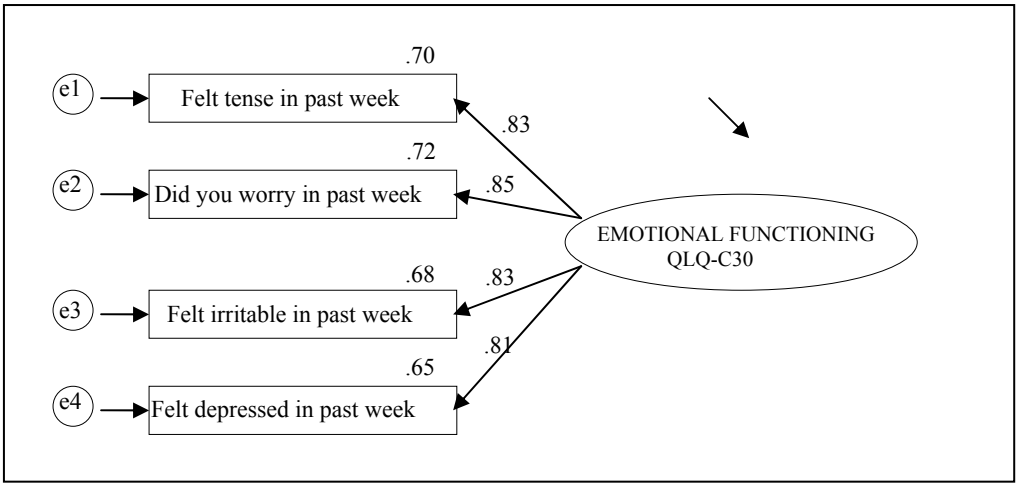
The results of the CFA for the QLQ-C30 physical functioning scale are presented in Figure 2. All five items loaded with critical ratios greater than two so are appropriate in the model. The standardized regression weights were above the 0.32 minimum (Tabachnick and Fidell 2001). The factor loadings for four of the five items were above 0.50, with three of these being above 0.75. Item five however only loaded at 0.39 (i.e. only 15% of the variance for this item was explained by the factor). Fit indices ranged from poor (RMSEA = 0.113), to borderline (AGFI = 0.88), to good (GFI = 0.96, CFI = 0.92). The internal consistency statistic (Cronbach's alpha) for the multi-item scale was good (α = 0.87). Cronbach's alpha would increase (α = 0.88) if item 5 was deleted.

Figure 2. Measurement model for the EORTC QLQ-C30 physical functioning scale showing parameter estimates and factor loadings for each item in the scale.



The results of the CFA for the EORTC QLQ-C30 emotional functioning scale are presented in Figure 3. All measures suggested an excellent fit. Factor loadings for the four component items ranged from 0.81 to 0.85 showing that 65% or more of the variance for each of the items was explained by the factor. Fit indices were excellent (GFI = 0.99, AGFI = 0.96, CFI = 1.00, RMSEA = 0.00). Cronbach's alpha for the multi-item scale was good (α = 0.90).

Figure 3. Measurement model for the EORTC QLQ-C30 emotional functioning scale showing parameter estimates and factor loadings for each item in the scale.



Cronbach's alpha for the two and three item scales ranged from acceptable (Cognitive functioning: α = 0.68; Nausea/vomiting: α = 0.67), to good (Social: α = 0.86; Pain: α = 0.87), to excellent (Role: α = 0.94; GHS: α = 0.94; Fatigue: α = 0.91).

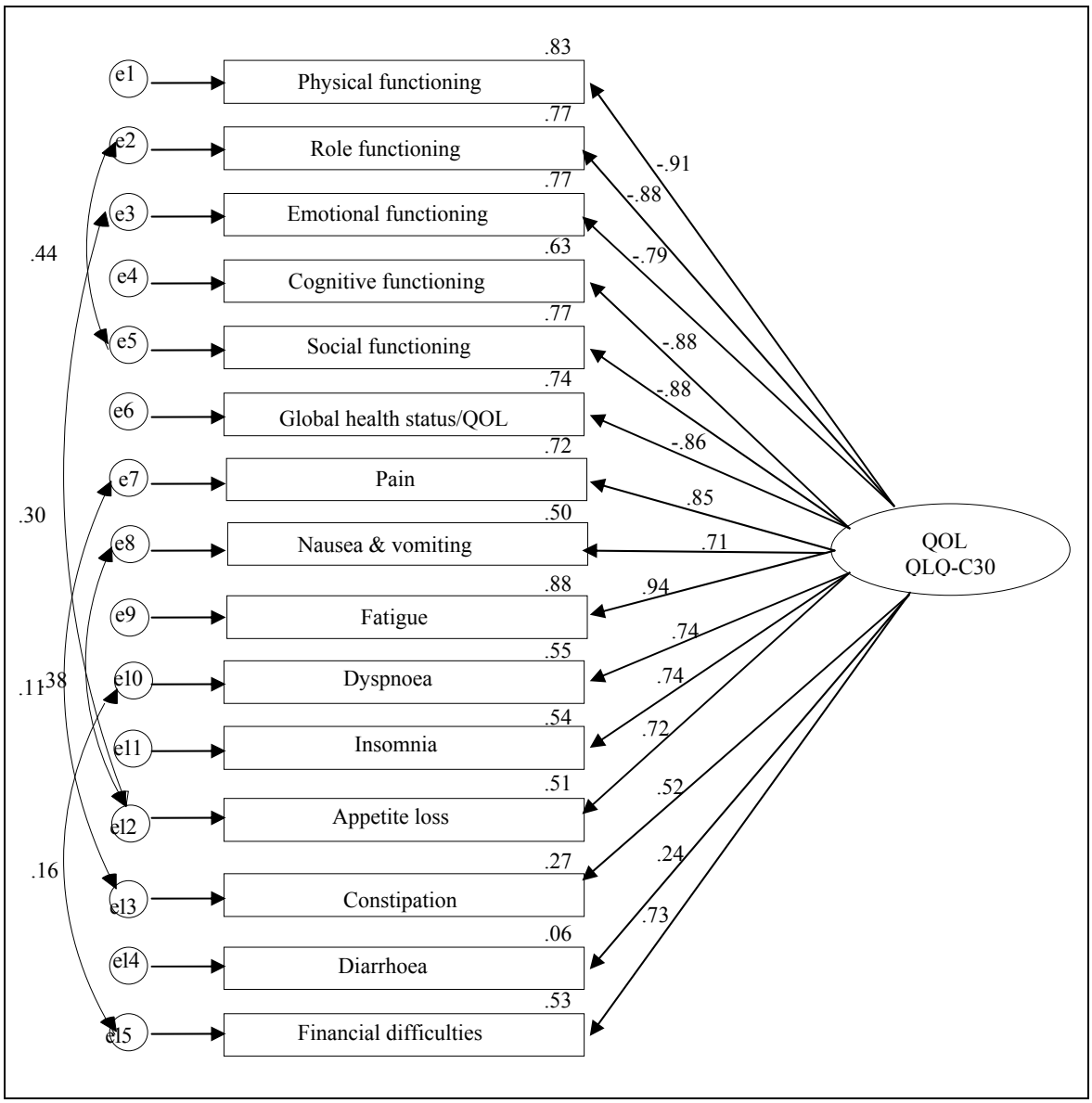
The factor loadings for the three component items of the EORTC QLC-LC13 ranged from 0.58 to 0.93. Cronbach's alpha was good (0.89).

The measurement models were then included in the structural models for the EORTC QLQ-C30 and the EORTC QLQ-LC13 and tested for fit.

Structural Models

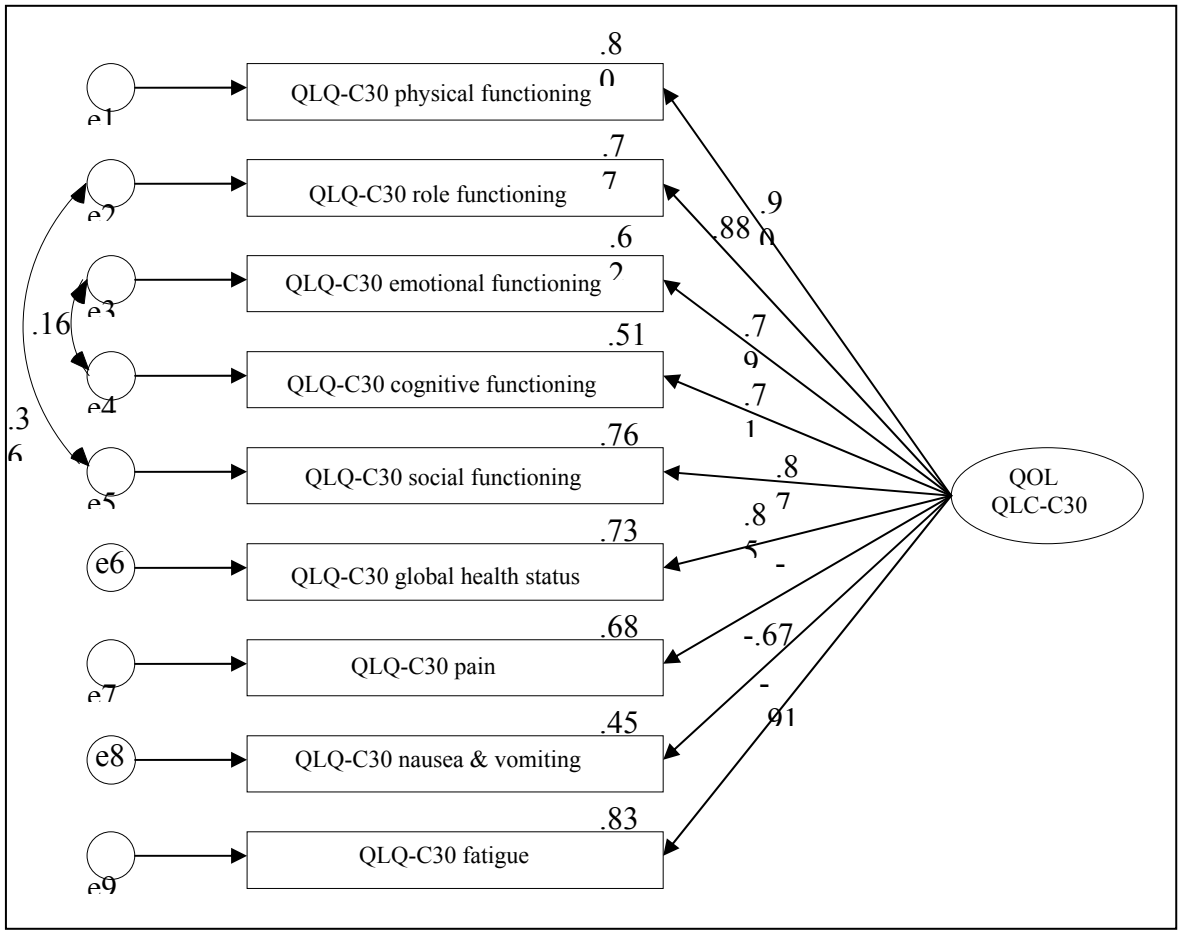
The initial EORTC QLQ-C30 structural model for one domain for quality of life (Figure 4) did not fit well. This model included the multi-item scales and the single items. It consisted of 5 functional scores, 1 global health score and 9 symptom scores. Initial analysis indicated that, for most of the items, factor loadings ranged from 0.71 to 0.94. Exceptions were constipation and diarrhea with factor loadings of 0.52 and 0.24 respectively. However, there were significant correlations between the error terms on: role functioning and social functioning (0.44); emotional functioning, appetite loss and nausea and vomiting (0.30 and 0.38); dyspnoea and financial difficulties (0.16); and pain and constipation (0.11). This usually indicates these relationships may represent another construct and should only be covaried if there is theory/evidence to substantiate the connection.

Figure 4. The initial EORTC QLQ-C30 structural model for one domain for QOL showing parameter estimates and factor loadings for each multi-item scale and single item, and significant correlations between the error terms.



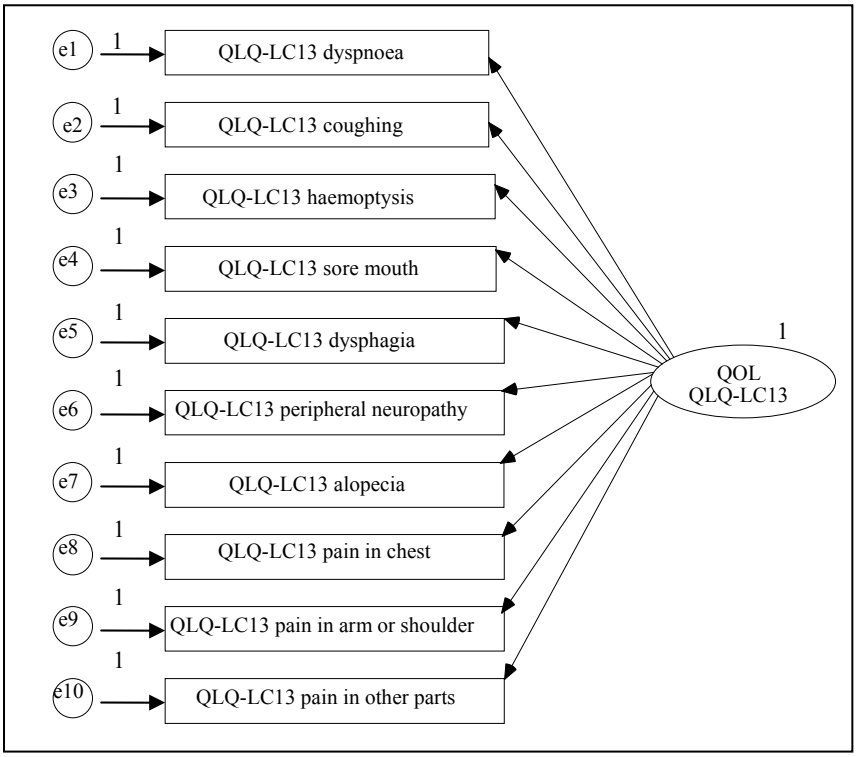
The model was then fitted to the multi item scores only (Figure 5). Results were mixed. Factor loadings ranged from 0.67 to 0.91. There were significant correlations between the error terms on role functioning and social functioning (0.36), and on emotional functioning and cognitive functioning (0.16). Fit indices were varied. Chi square was statistically significant ($\chi^2 = 38.6$, df = 25, p = 0.04) indicating the model did not fit well. Other indices were borderline (GFI = 0.94, AGFI = 0.89), to good (CFI = 0.91, RMSEA = 0.056).

Figure 5. The EORTC QLQ-C30 structural model fitted to multi-item scores only, showing parameter estimates and factor loadings for each multi-item scale, and significant correlations between the error terms.



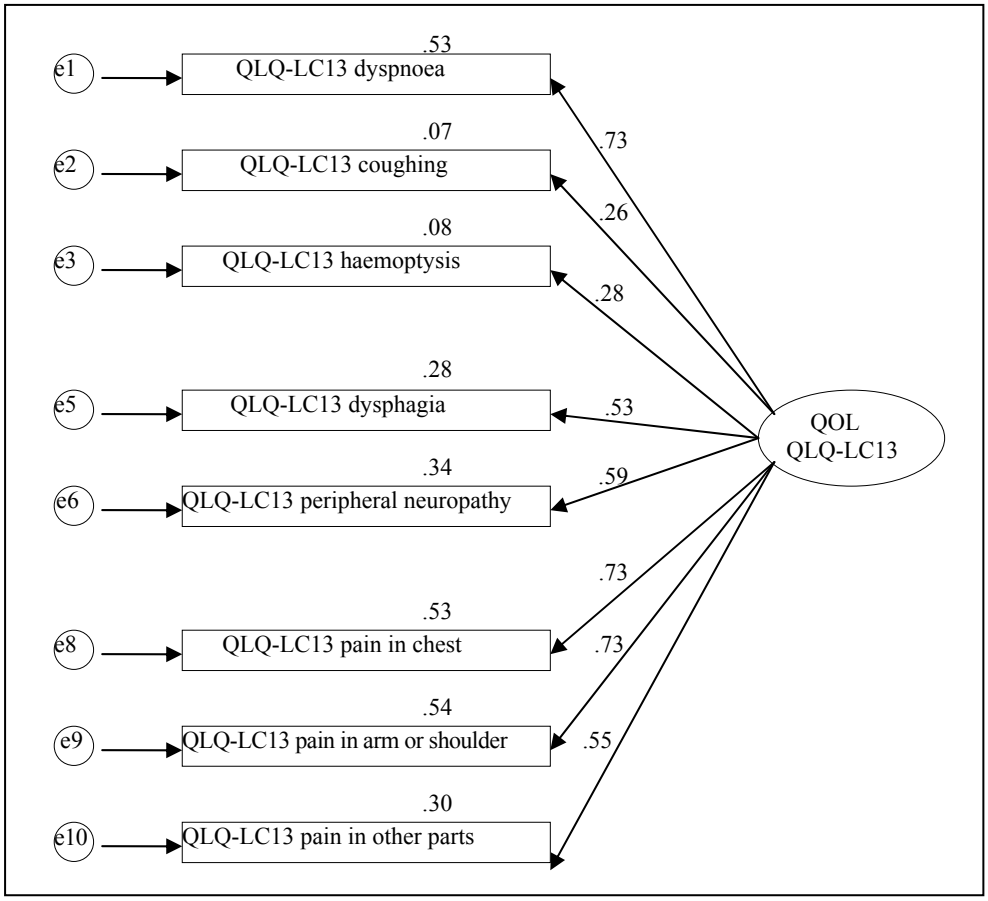
The initial EORTC QLQLC-13 structural (Figure 6) for one domain for quality of life did not fit well. The model consisted of 10 symptom scores. Several items (sore mouth, haemoptysis, peripheral neuropathy and alopecia) loaded with critical ratios less than 2. Chi square was statistically significant (χ^2 = 55.64, df = 35, p = 0.02) indicating the model did not fit well. Other indices were poor (CFI = 0.75), to borderline (GFI = 0.91, AGFI = 0.86), to good (RMSEA = 0.058).

Figure 6. The initial EORTC QLQ-LC13 structural model for one domain for QOL.



To improve the model, items 4 (sore mouth) and 7 (alopecia) were removed (Figure 7). Factor loadings then ranged from 0.53 to 0.73 except for coughing (0.26) and haemoptysis (0.28). Fit indices were varied. Chi square indicated an adequate fit ($\chi^2 = 26.3$, df = 20, p = 0.157). Other indices ranged from borderline (GFI = 0.95) to adequate (AGFI = 0.90, CFI = 0.90), to good (RMSEA = 0.043). No error terms covaried in this model.

Figure 7. The improved EORTC QLQ-LC13 structural model for one domain for QOL showing parameter estimates and factor loadings for the multi-item scale dyspnoea and single items (minus sore mouth and alopecia).



Convergent and Divergent validity

To test the convergent and divergent validity of the EORTC QLQ-C30 and the EORTC QLQ-LC13, correlation analysis was conducted on all multi-item scales and single items using the last time-point for each person. It was expected that there would be a moderate correlation among the physical-based functioning scales and the symptom scales, and among the psychosocial scales. The correlation among the physical and psychosocial scales was expected to be lower.

Results of the correlation analysis are reported in Tables 3-5. Correlations among the scales and single items of the QLQ-C30 (Table 3) were generally consistent with expectations (except between the physical and psychological scales), although they tended to be strong rather than moderate. More than 50% of the correlations among the physical-based functioning scales and the symptom scales were strong ranging from 0.50 to 0.81, and about 25% were moderate, ranging from 0.31 to 0.47. Among the psychological scales all correlations were strong (0.54 to 0.66), as were all correlations between the physical and psychological scales (0.52 to 0.80). Generally, correlations between the symptom single items and the multi-item function scales were the lowest with the majority being less than 0.50.

Table 3. Correlations among the EORTC QLQ-C30 scales and single items

	XPF	XRF	XEF	XCF	XSF	XQL	XPA	XNV	XFA	XDY	XSL	XAP	XCO	XDI	XFI
XPF	1.00														
XRF	0.81	1.00													
XEF	0.52	0.53	1.00												
XCF	0.58	0.63	0.55	1.00											
XSF	0.71	0.80	0.61	0.59	1.00										
XQL	0.76	0.71	0.56	0.54	0.66	1.00									
XPA	-0.71	-0.73	-0.53	-0.53	-0.64	-0.68	1.00								
XNV	-0.54	-0.55	-0.38	-0.37	-0.59	-0.52	0.50	1.00							
XFA	-0.80	-0.78	-0.65	-0.62	-0.72	-0.72	0.70	0.57	1.00						
XDY	-0.71	-0.62	-0.41	-0.45	-0.61	-0.55	0.46	0.40	0.60	1.00					
XSL	-0.51	-0.52	-0.52	-0.44	-0.50	-0.56	0.61	0.39	0.55	0.36	1.00				
XAP	-0.55	-0.54	-0.44	-0.41	-0.60	-0.49	0.46	0.56	0.59	0.42	0.47	1.00			
XCO	-0.50	-0.53	-0.49	-0.41	-0.55	-0.44	0.59	0.33	0.53	0.31	0.50	0.47	1.00		
XDI	-0.11	-0.08	-0.10	-0.10	-0.20	-0.12	0.01	0.20	0.13	0.12	0.13	0.19	0.03	1.00	
XFI	-0.34	-0.36	-0.47	-0.29	-0.44	-0.39	0.31	0.34	0.37	0.33	0.34	0.35	0.25	0.15	1.00

Approximate 95% confidence intervals are as follows: 0.90 (0.85, 0.93); 0.80 (0.72, 0.86); 0.70 (0.50, 0.79); 0.60 (0.46, 0.71); 0.50 (0.33, 0.64); 0.40 (0.22, 0.55); 0.30 (0.11, 0.47); 0.20 (0.00, 0.38).

Correlations among the EORTC-LC13 symptom scales and single items (Table 4) were fairly consistent with expectations, although there were slightly more small correlations than moderate correlations. This is consistent with the correlations between scales and single items of the QLQ-C30. About 42% were small (0.12 to 0.29), and 40% moderate (0.31 to 0.47).

Table 4. Correlations among the EORTC-LC13 scales and single items

	XLDY	XLCO	XLHA	XLSM	XLDS	XLPN	XLPO	XLAS	XLCH	XLAL
XLDY	1.00									
XLCO	0.41	1.00								
XLHA	0.29	0.25	1.00							
XLSM	0.46	0.18	0.31	1.00						
XLDS	0.47	0.40	0.15	0.55	1.00					
XLPN	0.35	0.07	0.00	0.31	0.21	1.00				
XLPO	0.35	0.23	0.18	0.28	0.32	0.21	1.00			
XLAS	0.35	0.01	0.14	0.28	0.12	0.42	0.37	1.00		
XLCH	0.54	0.25	0.2	0.36	0.41	0.36	0.39	0.47	1.00	
XLAL	-0.25	0.13	0.02	0.31	0.17	0.29	0.12	0.12	0.24	1.00

Approximate 95% confidence intervals are as follows: 0.90 (0.85, 0.93); 0.80 (0.72, 0.86); 0.70 (0.50, 0.79); 0.60 (0.46, 0.71); 0.50 (0.33, 0.64); 0.40 (0.22, 0.55); 0.30 (0.11, 0.47); 0.20 (0.00, 0.38).

Correlations among the scales and single items of the EORTC QLQ-C30 and QLQ-LC13 (Table 5) were generally consistent with expectations (except between the physical and psychological scales). Almost 50% of the correlations among the physical-based functioning scales and items, and the symptom scales and items were moderate ranging from 0.30 to 0.49. Correlations among the scales only were higher with about 50% being strong (0.53 to 0.82). Between the physical and psychological scales and items more than 60% of the correlations were moderate (0.30 to 0.49). Again correlations among the scales alone were higher (0.50 to 0.67).

Table 5. Correlations between the scales and single items in the EORTC QLQ-C30 and the EORTC-LC13

	XPF	XRF	XEF	XCF	XSF	XQL	XPA	XNV	XFA	XDY	XSL	XAP	XCO	XDI	XFI
XLDY	-0.77	-0.63	-0.50	-0.47	-0.67	-0.62	0.53	0.46	0.67	0.82	0.40	0.47	0.43	0.09	0.35
XLCO	-0.35	-0.32	-0.22	-0.13	-0.35	-0.35	0.24	0.30	0.33	0.37	0.19	0.35	0.23	0.04	0.21
XLHA	-0.31	-0.27	-0.22	-0.24	-0.28	-0.25	0.21	0.30	0.27	0.25	0.2	0.31	0.25	0.07	0.06
XLSM	-0.41	-0.39	-0.49	-0.39	-0.40	-0.36	0.35	0.38	0.40	0.37	0.29	0.33	0.25	0.06	0.26
XLDS	-0.45	-0.45	-0.41	-0.41	-0.47	-0.44	0.40	0.56	0.45	0.38	0.36	0.50	0.34	0.14	0.34
XLPN	-0.37	-0.35	-0.32	-0.30	-0.30	-0.34	0.33	0.12	0.35	0.30	0.24	0.08	0.18	-0.10	0.14
XLPO	-0.51	-0.50	-0.44	-0.45	-0.43	-0.51	0.63	0.40	0.47	0.27	0.49	0.32	0.42	0.10	0.21
XLAS	-0.38	-0.33	-0.39	-0.34	-0.35	-0.31	0.53	0.21	0.41	0.24	0.31	0.21	0.41	-0.05	0.27
XLCH	-0.54	-0.53	-0.45	-0.35	-0.56	-0.55	0.71	0.46	0.58	0.40	0.51	0.41	0.52	-0.02	0.33
XLAL	-0.28	-0.23	-0.25	-0.15	-0.23	-0.28	0.11	0.28	0.32	0.24	0.15	0.15	0.11	0.14	0.22

Approximate 95% confidence intervals are as follows: 0.90 (0.85, 0.93); 0.80 (0.72, 0.86); 0.70 (0.50, 0.79); 0.60 (0.46, 0.71); 0.50 (0.33, 0.64); 0.40 (0.22, 0.55); 0.30 (0.11, 0.47); 0.20 (0.00, 0.38).

Discriminant validity

Sensitivity to disease stage

The sensitivity of the scales to the large effects of moving from early (Stage I or II) disease to metastatic disease was tested with the HRQOL data at recruitment for the 113 patients whose disease did not progress contrasted with the HRQOL data from the last observation for the 45 patients whose disease did progress. Patients with newly diagnosed Stage I or II disease were expected to have much better HRQOL across all domains than patients with metastatic disease.

Differences in the mean scores for the EORTC QLQ-C30 and EORTC QLQ-LC13 domain scales for patients with Stage I or II disease and patients with metastatic disease are illustrated in the box and whisper plot (Figure 8). Mean differences and effect sizes are reported in Table 6. The mean differences for the scales were all in the expected direction, and, except for emotional functioning, were between 21.3 and 54.0, which is large relative to the scale range of 0-100. All but one of the effect sizes was greater than one, which is large according to conventional guidelines. The emotional functioning scale had the smallest mean difference and effect size. As the data was not normally distributed the non parametric Mann-Whitney U test was used to test the significance of differences for all domain scales and single items. All scores were significantly different (Table 6). For patients with metastases, functioning scores were all significantly lower and symptom scores were all significantly higher.

Figure 8. Box and whisper plot showing differences in the mean scores for the EORTC QLQ-C30 functioning and global health domain scales for patients with newly diagnosed Stage I or II disease (no metastases) at recruitment vs patients with metastatic disease at the last observation.

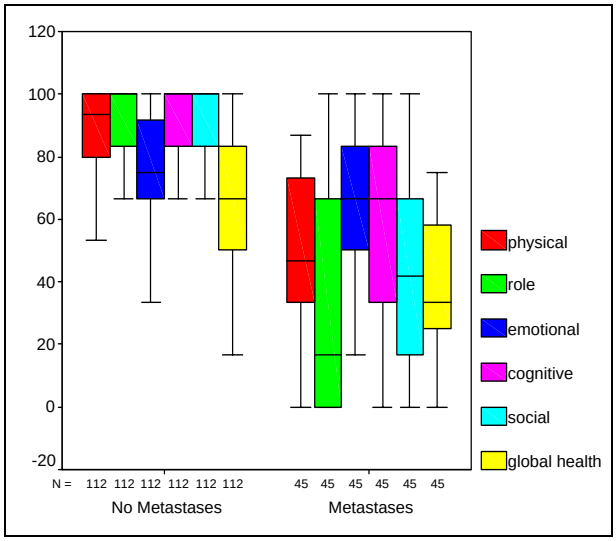


Table 6. Sensitivity of the EORTC QLQ-C30 and EORTC QLQ-LC13 domain scales to disease stage: Differences in mean scores, and effect sizes between patients with Stage I or II disease at recruitment (n = 112) vs patients with metastatic disease at the last observation (n = 45).

Scale	Mean difference (CI)[a]	Effect size
Physical functioning	38.3 (30.0, 46.7)***	1.89
Role functioning	54.0 (44.0, 64.1)***	1.97
Emotional functioning	13.7 (5.1, 22..4)**	0.60
Social functioning	45.3 (34.4, 56.2)***	1.72
Cognitive functioning	27.9 (18.1, 37.7)***	1.29
Global Health status/QOL	30.6 (22.9, 38.4)***	1.36
Fatigue	-40.7 (-50.0, -31.3)***	1.77
Nausea and Vomiting	-21.3 (-30.7, -12.0)***	1.14
Pain	-39.6 (-50.2, -28.9)***	1.52
Dyspnoea LC13	-35.6 (-45.0, -26.2)***	1.51

[a] Positive differences indicate patients with stage I or II disease score higher on functioning scales and lower on
symptom scales.
*P < 0.05; ** p < 0.01; ***p < 0.001

Sensitivity to differences between asymptomatic to mildly symptomatic

The sensitivity of the scales to the small effect of moving from asymptomatic to mildly symptomatic was tested with the HRQOL data at recruitment. Patients who were asymptomatic at recruitment (ECOG score of 0) were expected to have slightly better HRQOL across all domains than patients who were symptomatic at recruitment but whose symptoms had little or no impact on their daily function (ECOG score of 1).

Differences in the mean scores for the EORTC QLQ-C30 and EORTC QLQ-LC13 domain scales for patients with an ECOG score of 0 compared to those with an ECOG score of 1 are illustrated in the box and whisper plot (Figure 9). Mean differences and effect sizes are reported in Table 7. The mean differences (Table 7) for the scales were mostly in the expected direction (with the exception of the cognitive functioning and pain scales), and the largest was 7.2 points, which is small relative to the scale range of 0-100. The effect sizes (Table 7) ranged from 0.05 to 0.27, which is small according to conventional guidelines. The emotional functioning scale again had both the smallest mean difference and effect size.

Figure 9. Box and whisper plot showing the differences in mean scores for the EORTC QLQ-C30 domain scales at recruitment for patients with ECOG 0 vs 1 vs 2 or 3.

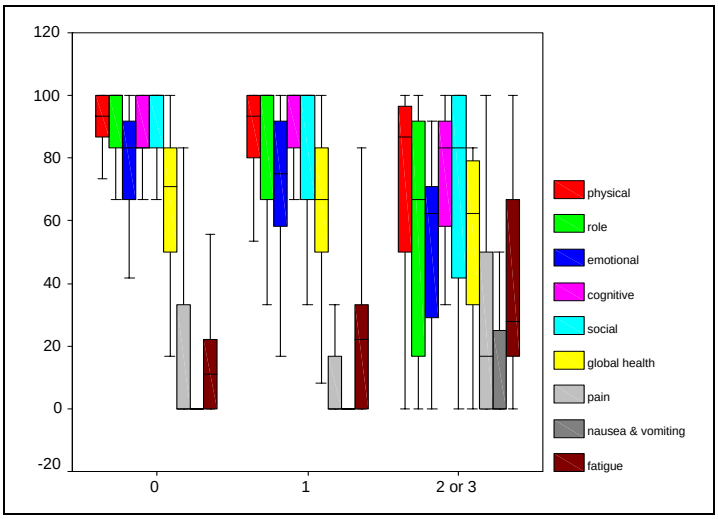


Table 7. Sensitivity of the EORTC QLQ-C30 and EORTC QLQ-LC13 domain scales to ECOG scores at recruitment: Differences in mean scores, and effect sizes between ECOG score 0 (n = 55-56 [a]) vs 1 (n = 106-107 [b])

Scale	**Mean difference (CI)[c]**	**Effect size**
Physical functioning [a, b]	3.9 (-1.3, 9.2)	0.24
Role functioning	7.2 (-1.6, 16.1)	0.25
Emotional functioning	1.2 (-7.0, 9.4)	0.05
Social functioning	5.8 (-2.4, 14.0)	0.22
Cognitive functioning	-3.7 (-10.3, 2.8)	0.20
Global Health status/QOL	5.6 (-2.3, 13.5)	0.23
Fatigue	-6.2 (-13.1, 0.6)	0.27
Nausea and Vomiting	-1.0 (-4.0, 2.0)	0.10
Pain	2.1 (-5.8, 10.0)	0.09
Dyspnoea LC13	-2.3 (-9.2, 4.7)	0.10

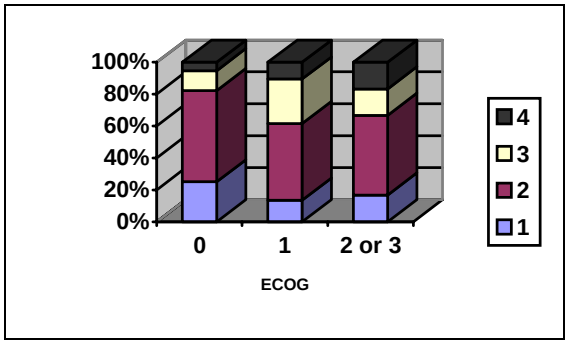
[a] Physical functioning (n=55), Remainder (n=56); [b] Physical functioning (n=107), Remainder (n=106)

[c] Positive differences indicate patients with an ECOG score of 0, score higher on functioning scales and lower on symptom scales.

*P < 0.05; ** p < 0.01; ***p < 0.001

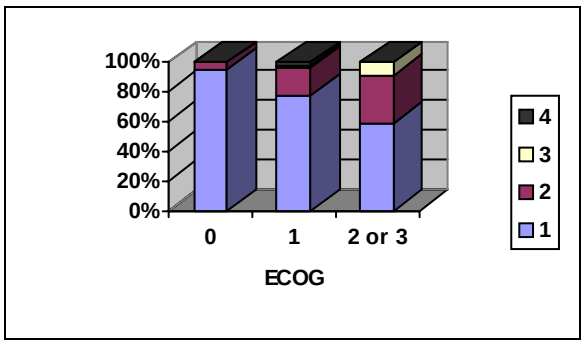
Differences in the mean scores for patients with ECOG scores of 0, 1, 2 and 3 were tested for statistical significance. As the data was not normally distributed the non parametric ANOVA (Kruskal Wallis test) was used. Significant differences were found only for the QLQ-C30 role functioning and emotional functioning scales (Fig 9), and the QLQ- LC13 single items coughing (Fig 10) and haemoptysis (Fig 11). Patients with an ECOG score of 2/3 had significantly lower role functioning (χ^2 = 9.036 (df, 2), p = 0.011), and emotional functioning (χ^2 = 7.219 (df, 2), p = 0.027) than patients with an ECOG score of 1 or 0. Patients with an ECOG score of 0 had significantly less coughing (χ^2 = 7.850 (df, 2), p = 0.020) and haemoptysis (χ^2 = 12.147 (df, 2), p = 0.002) than patients with an ECOG score of 1 or 2/3.

Figure 10. Distribution of the EORTC QLQ-LC13 QLQ-LC13 single single item coughing at recruitment for patients at recruitment for patients with ECOG scores 0 v 1 v 2/3 (be wary last grp very small)



1 = Not at all; 2 = A bit; 3 = Quite a bit; 4 = Very much

Figure 11. Distribution of EORTC item haemoptysis scores for ECOG 0 v 1 v 2/3



1 = Not at all; 2 = A bit; 3 = Quite a bit; 4 = Very much

Sensitivity to age

HRQOL data at recruitment was used to test the sensitivity of the domain scales to the small effect of age. It was expected that the HRQOL of older patients would be lower than for younger patients for all physical domains but not psychosocial domains;

Differences in mean scores for the EORTC QLQ-C30 and EORTC QLQ-LC13 domain scales for patients aged ≤ 60 years, 61-70 years and 70+ years, and effect sizes are reported in Table 8. The mean differences were mostly in the expected direction (with the exception of the symptom scales), but few were statistically significant. The largest difference was 14.3 (social functioning) which is moderate relative to the scale range of 0-100, all other differences were below 9 which is small. All but two of the effect sizes were less than 0.30, which is small according to conventional guidelines. The effect sizes for emotional functioning (0.35) and social functioning (0.46) were between small to modest by conventional standards. As the data was not normally distributed the non parametric ANOVA (Kruskal Wallis test) was used to test the significance of differences for all domain scales and single items. Only the QLQ-C30 social functioning scale and the single item, financial difficulties (Fig 13) discriminated between age groups. Social functioning was significantly lower for the younger 60 years and under age group (χ^2 = 10.412 (df, 2), p = 0.005), and financial difficulty was significantly higher (χ^2 = 10.476 (df, 2), p = 0.005). The box and whisper plot (Fig 12) is a further illustration of the differences for the social functioning and for the nausea and vomiting scales.

Table 8. Sensitivity of the EORTC QLQ-C30 and EORTC QLQ-LC13 domain scales to age at recruitment: Differences in mean scores, and effect sizes for ages ≤ 60 years (n=46-47 [a]) vs 60-70 years (n =59-63 [b]) and 60-70 years vs 70+ years(n = 62-64[c]).

	≤ 60 years vs 61 -70 years		**61-70 years vs 70+ years**	
Scale	**Mean difference (CI)[d]**	**Effect size**	**Mean difference (CI)[d]**	**Effect size**
Physical functioning	0.3 (-6.8, 7.4)	0.02	2.7 (-3.6, 9.0)	0.15
Role functioning	-5.9 (-18.4, 6.7)	0.18	-3.4 (-13.1, 6.2)	0.12
Emotional functioning	-1.5 (-11.5, 8.5)	0.06	-8.7 (-17.5, 0)	0.35
Social functioning	-14.3 (-26.5, -2.2)	0.46	-3.6 (-11.6, 4.4)	0.16
Cognitive functioning	5.5 (-1.8, 12.8)	0.27	-4.3 (-11.4, 2.8)	0.21
Global Health status/QOL	-2.8 (-12.8, 7.3)	0.10	-2.3 (-11.1, 6.5)	0.09
Fatigue	4.8 (-5.2, 14.9)	0.18	1.7 (-6.5, 9.8)	0.07
Nausea and Vomiting	2.9 (-1.5, 7.3)	0.24	0.3 (-3.6, 4.2)	0.03
Pain	0.7 (-9.5, 10.8)	0.02	4.7 (-3.9, 13.3)	0.19
Dyspnoea LC13	-1.7 (-10.7, 7.3)	0.07	4.7 (-3.5, 12.8)	0.20

[a] Pain (n=46); Remainder (n=47) [b] Pain (n=59); Remainder (n=63) [c] Pain (n=62); Remainder (n=64)

[d] Negative differences indicate that the function and HRQOL of older adults is worse than that of younger patients,
and that the symptom experience is better.

*P < 0.05; ** p < 0.01; ***p < 0.001

Figure 12. Box and whisper plot showing differences in the mean scores for the EORTC QLQ-C30 social functioning and nausea and vomiting domain scales at recruitment for age groups ≤60yrs (n = 47) vs 61-70yrs (n = 63) vs >70yrs (n = 64).

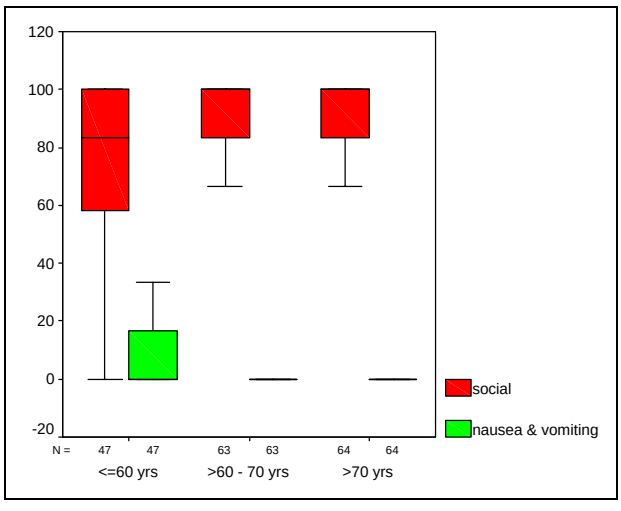
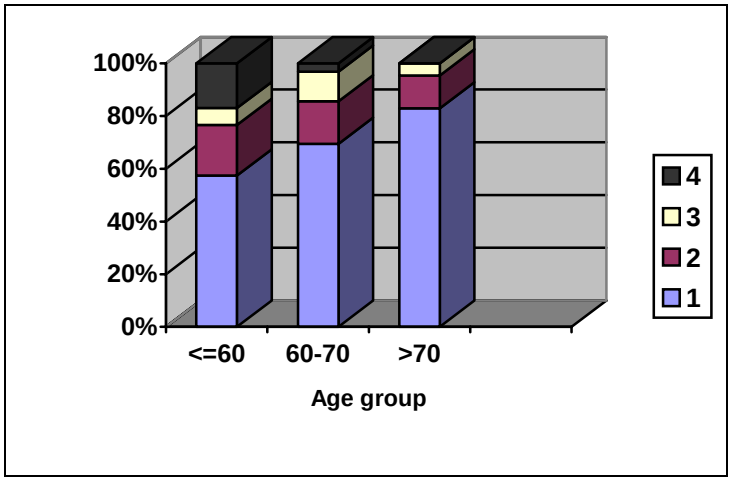


Figure 13. Distribution of the EORTC QLQ-C30 single item financial difficulties at recruitment by ≤60yrs (n = 47) vs 61-70yrs (n = 62) vs >70yrs (n = 64)



1 = Not at all; 2 = A bit; 3 = Quite a bit; 4 = Very much

Sensitivity to gender

The sensitivity of the scales to the small effect of gender was tested with the HRQOL data at recruitment. It was expected that men would have a slighter better HRQOL (about 5 points) than women.

The mean differences (Table 9) for the scales were mostly in the expected direction (with the exception of role functioning, pain and dyspnoea), and the largest was 7.1 points, which is small relative to the scale range of 0-100. All but one of the effect sizes (Table 9) was less than 0.28, which is small according to conventional guidelines. The cognitive functioning scale had the smallest mean difference and effect size. As the data was not normally distributed, the non parametric Mann-Whitney U test was used to test the significance of differences for all domain scales and single items. Only three scores discriminated between gender: males scored significantly higher on the QLQ-C30 emotional functioning scale (Table 9) ($z = -2.029$, $p = 0.042$) and significantly lower on the single item diarrhoea (Fig 16) ($z = -2.407$, $p = 0.016$); males scored significantly higher on the QLQ-LC13 single item coughing (Fig 15) ($z = -3.384$, $p = 0.001$). The box and whisper plot (Fig 14) also illustrates the extent of the differences for the emotional functioning and global health domain scales.

Table 9. Sensitivity of the EORTC QLQ-C30 and EORTC QLQ-LC13 domain scales to gender at recruitment: Differences in mean scores, and effect sizes between males (n = 124 -129 [a]) and females (n = 44-45 [b]).

Scale	Mean difference (CI)[c]	Effect size
Physical functioning	3.1 (-3.4, 9.6)	0.17
Role functioning	-1.9 (-12.2, 8.4)	0.06
Emotional functioning	7.1 (-1.1, 15.2)*	0.28
Social functioning	2.6(-7.4, 12.7)	0.09
Cognitive functioning	0.7 (-6.1, 7.4)	0.03
Global Health status/QOL	1.2 (-7.7, 10.1)	0.05
Fatigue	-0.9 (-9.4, 7.6)	0.04
Nausea and Vomiting	-1.9 (-6.0, 2.2)	0.17
Pain	3.0 (-4.2, 10.1)	0.12
Dyspnoea LC13	3.4 (-4.3, 11.2)	0.15

[a] Dyspnoea (n=124); Pain (n=128); Remainder (n=129) [b] Dyspnoea (n=44); Remainder (n=45)

[c] Positive differences indicate men score higher on functioning/QOL scales and lower on symptom scales.

*P < 0.05

Figure 14. Box and whisper plot showing differences in the mean scores for the EORTC QLQ-C30 emotional functioning and global health domain scales at recruitment by gender.

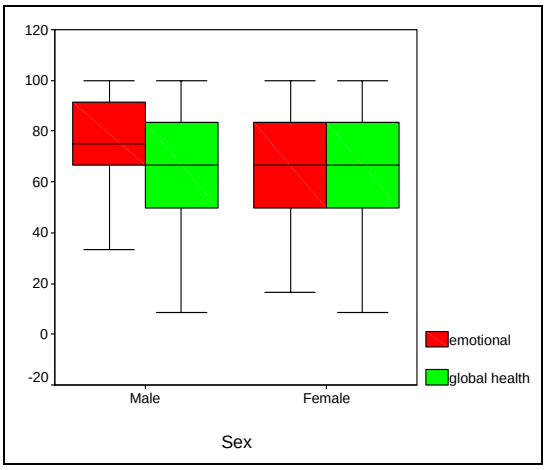
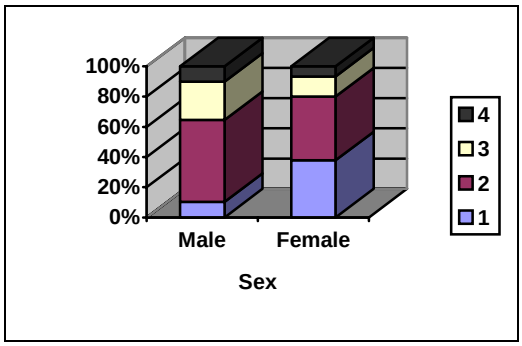
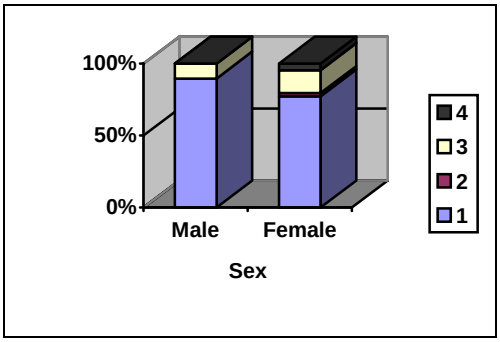


Figure 15. Distribution of the EORTC QLQ-LC13 single item coughing scores by gender



1 = Not at all; 2 = A bit; 3 = Quite a bit; 4 = Very much

Figure 16. Distribution of the EORTC QLQ-C30 single item diarrhoea scores by gender



1 = Not at all; 2 = A bit; 3 = Quite a bit; 4 = Very much

Sensitivity to comorbidities

The sensitivity of the scales to the small effect of comorbidities was tested with the HRQOL data at recruitment. It was expected that HRQOL would be inversely related to the number of comorbidities.

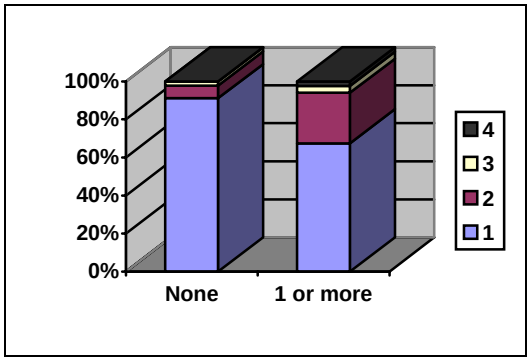
Differences in mean scores for the EORTC QLQ-C30 and EORTC QLQ- LC13 domain scales for patients with none and one or more comorbidities, and effect sizes are reported in Table 10. The mean differences for the domain scales were mostly in the expected direction, and the largest was 12.5 points, which is moderate relative to the scale range of 0-100. All but three of the effect sizes were less than 0.30, which is small according to conventional guidelines. The effect sizes for cognitive functioning (0.32) and fatigue (0.36) were between small to modest, and for dyspnoea (0.57), modest by conventional standards. As the data was not normally distributed, the non parametric Mann-Whitney U test was used to test the significance of differences for all domain scales and single items. Only seven scores discriminated between comorbidities: the QLQ-C30 physical functioning, cognitive functioning and fatigue scales (Table 10), and single items dyspnoea (Fig 18) and constipation (Fig 17); and the QLQ-LC13 dyspnoea scale (Table 10, Fig 20) and single item peripheral neuropathy (Fig 19). Patients with comorbidities had significantly worse physical ($z = -2.481$, $p = 0.013$) and cognitive ($z = -2.068$, $p = 0.039$) functioning, and significantly more fatigue ($z = -2.107$, $p = 0.035$), dyspnoea (QLQ-C30: $z = -2.448$, $p = 0.014$, QLQ-LC13: $z = -3.676$, $p = 0.000$), constipation ($z = -3.818$, $p = 0.000$) and peripheral neuropathy ($z = -2.039$, $p = 0.041$).

Table 10. Sensitivity of the EORTC QLQ-C30 and EORTC QLQ-LC13 domain scales to comorbidities at recruitment: Differences in mean scores, and effect sizes between patients with None (n = 86-90 [a]) vs 1 or more (n = 82-84[b]).

Scale	Mean difference (CI)[c]	Effect size
Physical functioning	5.3 (-0.1, 10.7)*	0.29
Role functioning	5.7 (-3.3, 14.6)	0.19
Emotional functioning	5.7 (-1.7, 13.2)	0.23
Social functioning	5.2 (-3.0, 13.4)	0.19
Cognitive functioning	6.3 (0.5, 12.2)*	0.32
Global Health status/QOL	6.0 (-1.5, 13.5)	0.24
Fatigue	-8.6 (-15.7, 1.4)*	0.36
Nausea and Vomiting	-2.5 (-5.8, 0.9)	0.22
Pain	-4.2 (-11.6, 3.2)	0.17
Dyspnoea LC13	-12.5 (-19.1, -5.9)**	0.57

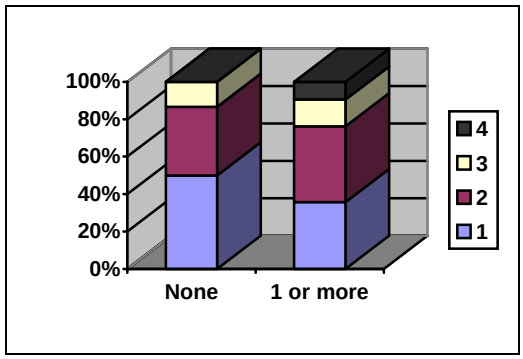
[a] Dyspnoea (n=86); Pain (n=89); Remainder (n=90) [b] Dyspnoea (n=82); Remainder (n=84)
[c] Positive differences indicate people with no comorbidities at recruitment score higher on functioning/QOL scales
and lower on symptom scales.
*P < 0.05; ** p < 0.001

Figure 17. Distribution of the EORTC QLQ-C30 single item constipation scores by comorbidities



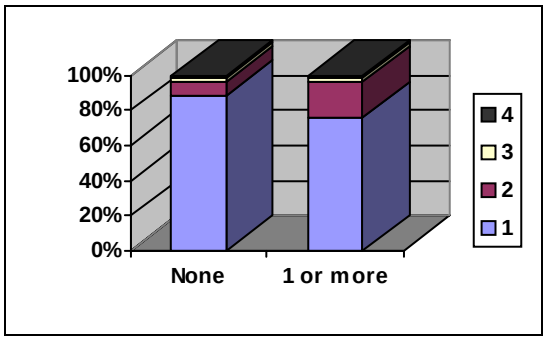
1 = Not at all; 2 = A bit; 3 = Quite a bit; 4 = Very much

Figure 18. Distribution of the EORTC QLQ-C30 single item dyspnoea scores by comorbidities



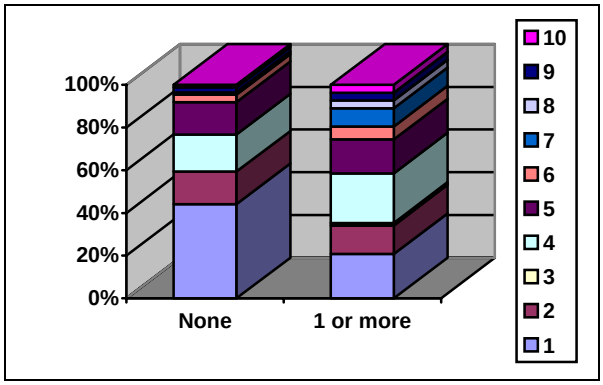
1 = Not at all; 2 = A bit; 3 = Quite a bit; 4 = Very much

Figure 19. Distribution of the EORTC QLQ-LC13 single item peripheral neuropathy scores by comorbidities



1 = Not at all; 2 = A bit; 3 = Quite a bit; 4 = Very much

Figure 20. Distribution of the EORTC QLQ-LC13 dyspnoea domain scale scores by comorbidities



Note: scale goes from 00 to 100 in 9 stteps

Reliability

Internal consistency and Test-re-test reliability

The reliability statistics for the EORTC QLQ-C30 and EORTC QLQ-LC13 domain scales are reported in Table 11. The Cronbach's alpha values for the domain scales were generally good to excellent, with the exception of the cognitive functioning and nausea scales. The intra-class correlation coefficients (ICC) were somewhat lower, but still generally acceptable to good (0.70 to 0.81) with the exception of the nausea (0.42) and pain (0.56) scales. Generally, this evidence supported the reliability of the scales of both instruments.

Table 11. Reliability of the EORTC QLQ-C30 and EORTC QLQ-LC13 domain scales: Internal consistency and test re-test statistics from HRQOL data taken at recruitment (n = 174) and recruitment vs admission to hospital (n = 132) respectively.

Scale	Cronbachs alpha	Intra-class coefficients
Physical functioning	.87	.81
Role functioning	.94	.74
Emotional functioning	.90	.78
Social functioning	.86	.70
Cognitive functioning	.68	.70
Global Health status/QOL	.94	.77
Fatigue	.91	.72
Nausea and Vomiting	.67	.42
Pain	.87	.56
Dyspnoea LC13	.89	.76

*P < 0.05; ** p < 0.01; ***p < 0.001

It is also useful to look at the mean change in the test-retest data (Table 12), as this suggests the degree of change that may be expected by chance for each scale. The mean differences ranged from 0.4 to -2.6 which is very small relative to the scale range of 0-100. All but two of the effect sizes were less than 0.10 which is very small according to conventional guidelines. A paired sample t-test, conducted to test the significance of the differences, showed that none were statistically significant.

Table 12. Test-retest reliability of the EORTC QLQ-C30 and EORTC QLQ-LC13 domain scales: Differences in mean scores, and effect sizes from recruitment to admission. (n = 126-133 [a,b]).

Scale	Mean difference (CI)	Effect size
Physical functioning	0.5 (-3.5, 4.6)	0.03
Role functioning	1.2 (-5.5, 7.9)	0.04
Emotional functioning	0.6 (-5.2, 6.4)	0.03
Social functioning	-2.6 (-8.7, 3.4)	0.10
Cognitive functioning	0.4 (-3.7, 4.6)	0.02
Global Health status/QOL	0.4 (-5.2, 6.0)	0.02
Fatigue	-1.1 (-6.2, 3.9)	0.05
Nausea and Vomiting	-1.0 (-3.2, 1.2)	0.11
Pain	0.8 (-4.2, 5.8)	0.04
Dyspnoea LC13	1.4 (-3.9, 6.7)	0.07

[a] Recruitment: Dyspnoea (n=126); Physical functioning (n = 132); Remainder (n=132)
[b] Admission: Dyspnoea (n=131), all others (n=133)

Responsiveness

Responsiveness to the effects of surgery

Responsiveness to the clinical effects of surgery was tested by comparing the HRQOL assessment scores at admission and discharge, of those who completed questionnaires at both these time points. A clinically large effect was expected, that is function symptoms and QOL were all expected to be worse at discharge, and this negative impact was expected to be large.

Differences in mean scores for the EORTC QLQ-C30 and EORTC QLQ- LC13 domain scales from admission to discharge, and effect sizes are reported in Table 13. The mean differences for the domain scales were all in the expected direction, and with the exception of emotional functioning, ranged from 14.5 to 53.9, which is large relative to the scale range of 0-100. All but one of the effect sizes were greater than 0.8, which is large according to conventional guidelines. The emotional functional scale had the smallest mean difference (5.8) and effect size (0.23). A paired sample t-test, conducted to test the significance of the differences, showed that all were statistically significant.

Table 13. Responsiveness of the EORTC QLQ-C30 and EORTC QLQ-LC13 domain scales to the effects of surgery: changes in mean scores, and effect sizes from admission to discharge (n = 122-137[a,b])

Scale	Mean difference (CI)[c]	Effect size
Physical functioning	-31.4 (-35.7, -27.0)***	1.51
Role functioning	-53.9 (-60.5, -47.3)***	1.77
Emotional functioning	-5.8 (-9.7, -1.9)**	0.23
Social functioning	-37.2 (-43.2, -31.2)***	1.20
Cognitive functioning	-19.6 (-23.7, -15.6)***	0.80
Global Health status/QOL	-26.6 (-31.2, -22.0)***	1.13
Fatigue	38.6 (33.8, 43.3)***	1.60
Nausea and Vomiting	14.5 (10.9, 18.2)***	0.83
Pain	47.5 (41.8, 53.2)***	1.76
Dyspnoea LC13	19.6 (14.5, 24.7)***	0.78

[a] Admission: Physical and Role functioning, Global health status, Pain, Dyspnoea (n=136); Remainder (n-137)

[b] Discharge: Role and emotional functioning, Global health status, Nausea and vomiting (n=136);
Dyspnoea (n-122); Remainder (n=137).

[c] Negative differences indicate deterioration in function/QOL over time and improvement in symptoms.

*P < 0.05; ** p < 0.01; ***p < 0.001

Responsiveness to disease progression

Responsiveness to disease progression was tested by comparing the HRQOL assessment scores at recruitment with those from the first and the last assessment following a diagnosis of disease recurrence, for the patients whose disease progressed within the study period and who completed questionnaires at each of these time points. A clinically large effect was expected, with functioning and QOL deteriorating and symptom experience increasing over time.

Differences in mean scores for the EORTC QLQ-C30 and EORTC QLQ-LC13 domain scales from recruitment to the first and last observation following recurrence, and effect sizes are reported in Table 14. When comparing the differences from recruitment to the first observation following recurrence, the mean differences for all the domain scales were in the expected direction, and all but one ranged from 15.8 to 31.5, which is large relative to the scale range of 0-100. Four of the effect sizes were greater than 0.8 which is large according to conventional guidelines. The remainder, with the exception of emotional functioning, ranged from 0.53 to 0.75 which is moderate. The emotional functioning scale was the only one with a small mean difference (4.6) and effect size (0.15). A paired sample t-test, conducted to test the significance of the differences, showed that, with the exception of emotional functioning, all were statistically significant. The differences and effect sizes continued to grow over time. Comparing recruitment to the last observation following a diagnosis of recurrence, the differences were all larger than at the first observation following metastases. All of the effect sizes except for emotional functioning were now large (over 0.8). The effect size for emotional functioning was still small (0.19). Paired sample t-tests showed that all differences, except for emotional functioning were statistically significant.

Table 14. Responsiveness of the EORTC QLQ-C30 and EORTC QLQ-LC13 domain scales to the effects of disease progression: changes in mean scores, and effect sizes from recruitment to first and last observation after metastases (n = 33-37 [a, b])

	Recruitment to first observation following metastases		Recruitment to last observation following metastases	
Scale	**Mean difference (CI) [c]**	**Effect size**	**Mean difference (CI) [c]**	**Effect size**
Physical functioning	-30.7 (-40.2, -21.1)***	1.40	-41.4 (-51.5, -31.4)***	1.89
Role functioning	-31.5 (-44.3, -18.8)***	0.97	-41.0 (-54.9, -27.1)***	1.28
Emotional functioning	-4.6 (-14.9, 5.7)	0.15	-5.5 (-16.2, 5.3)	0.19
Social functioning	-23.0 (-35.3, -10.7)**	0.67	-32.0 (-43.9, -20.0)***	0.96
Cognitive functioning	-15.8 (-28.5, -3.1)**	0.53	-25.2 (-36.2, -14.2)***	0.82
Global Health status/QOL	-18.5 (-28.2, -8.8)***	0.67	-21.6 (-32.0, -11.3)***	0.84
Fatigue	22.2 (11.9, 32.5)***	0.75	33.9 (23.9, 44.0)***	1.17
Nausea and Vomiting	20.3 (10.5, 30.1)***	0.86	23.4 (13.2, 33.6)***	0.94
Pain	31.5 (21.1, 41.8)***	1.10	39.3 (28.1, 50.6)***	1.33
Dyspnoea LC13	20.4 (9.7, 31.1)***	0.72		
Dyspnoea LC13			31.6 (20.8, 42.4)***	1.11

[a] Recruitment and First observation: Dyspnoea (n=33); Remainder (n=37)
[b] Last observation: Dyspnoea (n=35); Remainder (n=37)
[c] Negative differences indicate deterioration in function/QOL over time and improvement in symptoms.
*p < 0.05; ** p < 0.01; ***p < 0.001;
All scales except for emotional functioning are significant from recruitment to last observation
*** p < 0.001

Responsiveness to the effects of radiotherapy

Responsiveness to the clinical effects of radiotherapy was tested by comparing the HRQOL assessment scores at recruitment with those from the first and the last radiotherapy treatment, for the patients who had radiotherapy and who completed questionnaires at each of these time points. A clinically moderate to large effect was expected in the short term, with radiotherapy having a deleterious impact on all domains of QOL. Differences in mean scores for the EORTC QLQ-C30 and EORTC QLQ-LC13 domain scales from recruitment to the first and last radiotherapy treatments, and effect sizes are reported in Table 15.

Comparing the differences from recruitment to the first radiotherapy treatment, the mean differences for the domain scales were mostly in the expected direction (except for emotional and cognitive functioning). The majority of the differences were moderate (9.1 to 13.8 and 15.3 for role functioning) relative to the scale range of 0-100. Cognitive functioning, global health nausea and vomiting were trivial to small (1.8 to 4.6). Only two of the effect sizes were large (physical and role functioning at 0.99 and 1.01), the remainder were small to moderate (0.09 to 0.67) according to conventional guidelines. Cognitive functioning had the smallest mean difference (1.8) and effect size (0.09). A paired sample t-test, conducted to test the significance of the differences, showed that physical, role and social functioning, fatigue, pain and dyspnoea were statistically significant.

Comparing the first to last radiotherapy treatment, differences were again mostly in the expected direction and most with a further small to moderate (6.5 to 9.3) difference. Most of the effect sizes were small (0.19 to 0.39). Physical functioning had the smallest mean difference (0.02).and effect size (0.00). A paired sample t-test, conducted to test the significance of the differences, showed that social functioning, global health status and fatigue were statistically significant.

Differences from recruitment to the last radiotherapy treatment were all in the expected direction and mostly moderate to large (8.8 to 22.2). Effect sizes were mostly moderate (0.49 to 0.74) to large (0.87). The effect sizes for emotional and cognitive functioning and global health were only trivial or small. Cognitive functioning again had the smallest mean difference and effect size. Significance testing showed that, as for recruitment to first radiotherapy treatment, the mean differences for physical, role and social functioning and fatigue, pain and dyspnoea were significantly different.

Table 15. Responsiveness of the EORTC QLQ-C30 and EORTC QLQ-LC13 domain scales to the effects of adjuvant radiotherapy: changes in mean score from Recruitment to First and Last radiotherapy treatments (n = 30-36[a])

	Recruitment to first radiotherapy treatment		**First to last radiotherapy treatment**		**Recruitment to last radiotherapy treatment**	
Scale	**Mean difference (CI)[c]**	**Effect size**	**Mean difference (CI)[c]**	**Effect size**	**Mean difference (CI)[c]**	**Effect size**
Physical functioning	-13.8 (-18.9, -8.7)***	0.99	0.02 (-5.6, 5.7)	0.00	-13.8 (-20.5, -7.1)***	0.87
Role functioning	-15.3 (-27.8, -2.7)*	1.01	-6.9 (-16.4, 2.5)	0.22	-22.2 (-36.2, -8.3)*	0.72
Emotional functioning	9.1 (-0.2, 18.4)	0.37	-1.8 (-5.5, 1.8)	0.08	7.3 (-3.1, 17.6)	0.28
Social functioning	-11.1 (-21.5, -0.8)*	0.45	-9.3 (-17.9, -0.6)*	0.32	-20.4 (-33.2, -7.6)*	0.69
Cognitive functioning	1.8 (-7.6, 11.3)	0.09	-3.2 (-7.2, 0.8,)	0.19	-1.4 (-11.9, 9.2)	0.06
Global Health status/QOL	-4.6 (-13.8, 4.5)	0.21	-6.5 (-12.2, -0.8)*	0.32	-11.1 (-21.0, -1.2)	0.49
Fatigue	9.9 (1.3, 18.4)*	0.47	6.6 (0.4, 12.8)*	0.31	16.5 (6.7, 26.3)*	0.73
Nausea and Vomiting	2.3 (-2.6, 7.2)	0.21	6.5 (-0.3, 13.3)	0.39	8.8 (-1.9, 15.7)	0.57
Pain	13.3 (4.2, 22.4)**	0.67	1.9 (-5.1, 8.8)	0.09	15.2 (5.4, 25.1)**	0.74
Dyspnoea LC13	10.0 (3.7, 16.3)**	0.53	4.1 (-2.6, 10.7)	0.19	14.1 (7.4, 20.7)***	0.68

[a] Dyspnoea (n=30); Pain (n=35); Remainder (n=36)

[c] Negative differences indicate deterioration in function/QOL over time and improvement in symptoms

*P < 0.05; ** p < 0.01; ***p < 0.001

Discussion

Health related quality of life (HRQOL) is an important component of the evaluation of treatment, particularly in chronic diseases where the aim of treatment is to ameliorate symptoms rather than extend survival. The assessment of HRQOL in cancer care is particularly important because the physical and psychological effects of the disease, and the benefits and toxicities of cancer treatments have a direct effect on patients' wellbeing and their ability to perform their usual roles and abilities (Schipper and Clinch 1988; Fallowfield 1990; Cella and Tulsky 1993b; Cleton 1995; Greco 1995; Hanks and Hoskin 1995; Maguire 1995; Steel 1995).

The Quality of Life Questionnaire Core module (QLQ-C30) (Aaronson, Ahmedzai et al. 1993) is the most widely used instrument for measuring HRQOL in clinical trials. It is the core component of the European Organisation for Research and Treatment of Cancer (EORTC)'s modular approach to QOL assessment and represents QOL domains relevant across a wide range of cancer sites and treatment types. The QLQ-C30 is complemented by modules specific to particular cancers such as the lung cancer module (QLC-LC13) which is meant for use with a wide variety of lung cancer patients in varying disease stage and treatment modality (Bergman, Aaronson et al. 1994). Both the QLQ-C30 and the QLQ-LC13 have been previously validated. A bibliography of validation studies for the QLQ-C30 can be found in the EORTC QLQ-C30 scoring manual (Fayers, Aaronson et al. 2001) and a summary of findings from numerous studies can be found in Spilker et al (Spilker 1996). The QLQ-LC13 was validated in field tests together with the QLQ-C30 (Bergman, Aaronson et al. 1994), and has subsequently been validated in other studies (Chie, Yang et al. 2004; Nowak, Stockler et al. 2004).

However the validity of a HRQOL instrument is not something that is established by a single or even a few studies. Whether or not the instrument produces sensible and useful results in various circumstances should be judged in an ongoing process of validation (Streiner and Norman 1996; Fayers and Machin 2000). Although the QLQ-C30 and the QLQ-LC13 have undergone a continual process of validation across a range of health care contexts and disease groups, and in different nationalities and cultures, our confidence in, and understanding of, the instruments will develop as the body of evidence accrues. The results from this study contribute to the growing literature supporting the scientific validity, reliability and responsiveness of the QLQ-C30 and QLQ-LC13 as effective measures of HRQOL in the area of cancer care, and, more specifically, as measures of HRQOL in an Australian sample of people with early stage non-small cell lung cancer (NSCLC).

The factor structure reported previously for the QLQ-C30 (Aaronson, Ahmedzai et al. 1993) and the QLQ-LC13 (Bergman, Aaronson et al. 1994) was generally replicated in this sample, confirming the questionnaires' construct validity. Confirmatory factor analysis (CFA) conducted on the physical functioning scale of the QLQ-C30 showed that all measures were either adequate or better. Factor loadings for four of the five items ranged from 0.51 to 0.95, goodness of fit indices were adequate and internal consistency was good (0.88). CFA on the emotional functioning scale suggested an excellent fit. Factor loadings were all above 0.80, goodness of fit indices were excellent as was internal consistency (0.90). Internal consistency for the two and three item scales ranged from acceptable to excellent. For the QLQ-LC13 dyspnoea scale, factor loadings were acceptable (0.58 – 0.93) and internal consistency was good (0.89). When these measurement models were included in the structural models our findings were however mixed. For the QLQ-C30 model (multi-items scores only), factor loadings were good (0.67 to 0.91) but fit indices varied from poor to good. For the QLQ-LC-13 model

(without sore mouth and alopecia which were taken out to improve the model) factor loadings varied from poor to good (0.26 – 0.73) but indices of fit measures were generally adequate. However, consideration of clinical validity probably outweigh such psychometric concerns in this context. Given that the QLQ-C30 is well established and widely used in its current form, the results of this study are unlikely to lead to modifications of the instrument.

Our study also confirmed the convergent and divergent validity of the instruments. We expected moderate correlations among the physical-based functioning scales and symptoms scales, and among the psychosocial scales. The results were generally in agreement with our expectations although correlations tended to be strong rather than moderate. For the QLQ-C30, more than 50% of the correlations among the physical-based functioning scales and symptom scales were strong (0.50 to 0.81) and approximately 25% were moderate (0.31 to 0.47). All correlations among the psychosocial scales were strong (0.54 to 0.66). This is in line with numerous other studies that also found strong correlations among these scales, particularly between the physical, role fatigue and pain scales. However, contrary to expectations we found strong correlations among the physical and psychosocial scales (0.52 to 0.80). While most previous studies found small correlations among these scales, there were several studies that had similar results to ours. Some (Osoba, Zee et al. 1994; Blazeby, Conroy et al. 2003; Cocks, Cohen et al. 2007) found moderate to large correlations between the emotional functioning and physical functioning, role functioning and nausea and vomiting scales. Moderate to large correlations between the emotional functioning and fatigue and pain scales were also found (Aaronson, Ahmedzai et al. 1993; Osoba, Zee et al. 1994; Kaasa, Bjordal et al. 1995; Fitzsimmons, Kahl et al. 2005; Cocks, Cohen et al. 2007). While correlations among the QLQ-LC13 symptom scales and single items were also generally consistent with our expectation we did find slightly more small (42%) than moderate (40%) correlations.

Correlations among scales and single items of the QLQ-C30 and QLQ-LC13 were also generally consistent with expectations (50% of correlations between scales and items were moderate and 50% correlations between scales only were strong). A strong correlation between the dsypnoea scale on the QLQ-LC13 and the dyspnoea item on the the QLQ-C30 was also found in a previous study (Chie, Chang et al. 2003). However, as for correlations within each of the individual instruments, correlations between the physical and psychological scales and items of the two instruments were contrary to our expectations with more than 60% being moderate and those among the scales even higher (0.50 to 0.67).

The sensitivity of the scales to the expected effects of progressive disease, age, gender, and comorbidities, provide further support for the questionnaires' discriminant validity. In terms of moving to metastatic disease, differences in the QLQ-C30 scale scores were all in the expected direction and, except for emotional functioning, were large (21.3 to 54.0) and significant, as were the effect sizes (1.14 to 1.97). Although the effect size for emotional functioning was moderate (0.60), the differences were statistically significant. These results generally confirm findings from other studies evaluating differences in quality of life for patients with local disease compared to those with metastatic disease (Aaronson, Ahmedzai et al. 1993; Osoba, Zee et al. 1994). In relation to the QLQ-LC13 however, our findings were contrary to previous studies that found no significant difference in scores on the dyspnoea scale for local and metastatic disease patients (Bergman, Aaronson et al. 1994; Nicklasson and Bergman 2007). They did however find significant differences for the single items coughing and haemoptysis which we also found.

In terms differences between asymptomatic (ECOG 0) and symptomatic but completely ambulant (ECOG 1) disease, our study revealed small differences in scores and effect sizes as expected, indicating that both instruments are sensitive to this small effect. For the QLQ-C30, the largest difference was 7.2 for role functioning and the smallest was 1.2 for emotional functioning. Differences for the cognitive functioning and pain scales were however not in the expected direction. This is understandable for cognitive functioning, as it is not easily observed (by the person rating the ECOG status of the patient) and is not usually considered a symptom per se. It is less understandable for pain, since this is a common symptom in cancer, but perhaps not in fully ambulant lung cancer patients. The mean difference for the QLQ-LC13 dyspnoea scale was -2.3. It is difficult to compare our results here to other studies as most either compare differences across all, or combinations of, ECOG levels (Bergman, Aaronson et al. 1994; Osoba, Zee et al. 1994; McLachlan, Devins et al. 1998; Brabo, Paschoal et al. 2006), or use different criteria to measure performance status such as the World Health Organisation WHO (Nicklasson and Bergman 2007) or Karnoffsky Performance Status (KPS) (Bjordal, de Graeff et al. 2000). These studies did however find the instruments were able to detect a difference in performance status by disease stage. Our study showed that they were not only able to detect differences, but to detect differences in the smallest change in disease status, namely moving from asymptomatic (ECOG 0) to mildly symptomatic (ECOG 1) disease.

Our findings in relation to age, gender and comorbidity reflect findings from other studies, though these are limited. For the QLQ-C30 we found that for several scale, mean scores deteriorated as age increased. Most of these differences were small to moderate, as expected, except for social functioning which was large, especially for ages under 60 years compared to 71-70 years. Role functioning had the second largest difference although this was classified as a moderate change. Lund Hagelein et al (2006) found the same trend and, as in our study, the largest differences were in the social functioning scale followed by role functioning. They however, found that scores on the emotional functioning scale improved with age which was contrary to our findings. Contrary to expectation, our study found that scores on the symptom scales improved slightly (though not significantly) with age. This trend was also found to some degree in the study by Lund Hagelin et al (2006). It is not clear why this should be the case, but perhaps older patients are more accepting of poorer health generally due ot the loss of health that occurs with aging, and that this enhances their ability to cope and therefore reduces their perception of the degree of symptoms, which in turn influences their self-report of their symptom experience.

The differences we found in scores for gender were small as expected and mostly in the expected direction. Mean differences and effect sizes were small (largest difference was 7.1 for QLQ-C30 emotional functioning and smallest was 0.7 for QLQ-C30 cognitive functioning). Only a few differences were statistically significant. Males scored significantly higher on the QLQ-C30 emotional functioning scale and significantly lower on the single item diarrhoea; males scored significantly higher on the QLQ-LC13 single item coughing. Small though non significant differences in scale scores for the QLQ-C30 were also found in a sample of patients with pancreatic cancer (Fitzsimmons, Kahl et al. 2005) and also in a study of terminally ill cancer patients (Lundh Hagelin, Seiger et al. 2006). That study however found that women had significantly higher scores for nausea and vomiting whereas in our study scores were higher for women but they were not significant.

In terms of sensitivity to comorbidities, mean differences for the domain scales were mostly in the expected direction, and were small to moderate. The largest was for the QLQ-LC13,

Dyspnoea, which had a difference of 12.5 points. Differences for the QLQ-C30 physical functioning, cognitive functioning and fatigue scales and single items constipation and dyspnoea, and the QLQ-LC13 dyspnoea scale and single item peripheral neuropathy, were statistically significant. We only found one other study ((Michelson, Bolund et al. 2000) that had investigated the instruments sensitivity in relation to comorbidities. The sample in this study did not consist of cancer patients but rather a random sample of individuals with chronic health problems including cancer. Their results also showed that the QLQ-C30 was able to discriminate between people with differing levels of comorbidities.

When evaluating the sensitivity of the instruments to the small effects of moving from asymptomatic to mildly symptomatic disease, age, gender and comorbidities it should be remembered that while differences are small and may not be significantly different these differences may still be clinically relevant (King 1996). For the QLQ-C30 a difference of 5 points (or 10 for role functioning) is relatively small but may be clinically important. For the emotional and cognitive functioning scales and for the QLQ-LC13 dyspnoea scale, there is insufficient evidence to date to judge the amount of difference required for clinical relevance. In our study all small effects were detected by the QLQ-C30 and the QLQ-LC13. Although the differences were small, in many cases they were clinically relevant.

The reliability of both instruments was also confirmed in this sample. The Cronbach's alpha values for the domain scales of both the QLQ-C30 and QLQ-LC13 were generally excellent (0.86 to 0.94), with the exception of the cognitive functioning (0.68) and nausea (0.67) scales. For the QLQ-C30, our cronbachs alphas were generally higher than those in earlier validation studies but similar to those in more recent studies that used the revised version (version 3)(Aaronson, Ahmedzai et al. 1993; Ringdal and Ringdal 1993; Fossa 1994) (Osoba, Zee et al. 1994; Kaasa, Bjordal et al. 1995; Ringdal, Ringdal et al. 1999; Bjordal, de Graeff et al. 2000; Martin, Rubenstein et al. 2003; Nowak, Stockler et al. 2004) (Galalae, Loch et al. 2004; Fitzsimmons, Kahl et al. 2005; Galalae, Michel et al. 2005; Luo, Fones et al. 2005; Poveda, Lopez-Pousa et al. 2005; Boehmer and Luszczynska 2006; Easson, Bezjak et al. 2007) (see Appendix, Table A1). The cognitive functioning scale, as it did in all previous studies, had the lowest cronbach's alpha. Our role functioning scale had the highest cronbach's alpha. This is not surprising as despite earlier studies reporting low alphas for this scale, more recent studies, using version 3, all report high alphas. The cronbach's alpha for the QLQ-LC13 is also similar to those reported in previous studies (Bergman, Aaronson et al. 1994; Chie, Yang et al. 2004; Brabo, Paschoal et al. 2006; Nicklasson and Bergman 2007) (see appendix) . The intra-class correlation coefficients (ICC) were somewhat lower than the cronbach's alphas, but still generally acceptable to good (0.70 to 0.81) with the exception of the nausea (0.42) and pain (0.56) scales. Low ICC's for the nausea and pain scales were also found in previous studies (Martin, Rubenstein et al. 2003; Chie, Yang et al. 2004). The distributions of the nausea and pain scales are usually skewed ,with the majority of scores at the low end of the scale (floor effects) (Fayers, Weeden et al. 1998). The low ICC scores for these scales are as much a refection o the low variability in scores among people as large mean change over a period when change should not occur (ie, poor reliability). This is one of the limitations of the ICC as a measure of reliability. This is why we also assessed reliability through the mean differences from the test-retest data. Differences ranged from 0.4 to -2.6, which is very small and t-tests showed that none of the differences were statistically significant, confirming the reliability of all the scales.

The responsiveness of most scales of the QLQ-C30 and the QLQ-LC13 to the large effects of surgery and radiotherapy and to disease recurrence support the instruments' longitudinal validity

in this sample. In terms of surgery, with the exception of the emotional functioning scale (5.8, 0.23), mean differences and effect sizes for all QLQ-C30 scales were large (14.5, 0.83 for nausea and vomiting to 53.9, 1.77 for role functioning) and significant. These large differences are similar to those found in studies of patients with colorectal liver metastases who underwent surgery (Krabbe, Peerenboom et al. 2004). Other studies of patients with head and neck cancer (Bjordal, de Graeff et al. 2000) and prostate cancer (Krahn, Bremner et al. 2007) found small to moderate differences as a result of treatment. In these two studies treatment types were combined and included surgery, radiotherapy, hormone therapy, chemotherapy, or combinations of these. The dyspnoea scale of the QLQ-LC13 was also found to be responsive to the effects of surgery with a mean difference and effects size of 19.6 and 0.78.

In relation to radiotherapy, in this sample, mean differences from recruitment to the first radiotherapy were mostly moderate (9.1 to 13.8 and 15.3 for role functioning;) and effect sizes ranged mostly from small to moderate (0.09 to 0.67) with only physical and role functioning at 0.99 and 1.01 being large. By the last radiotherapy treatment, mean differences were mostly moderate to large (8.8 to 22.2) and effect sizes were mostly moderate to large (0.49 to 0.87). Significance testing showed that differences for the QLQ-C30 physical, role and social functioning, fatigue, and pain scales and the QLQ-LC13 dyspnoea scale were statistically significant, that is, patients experienced significant deterioration in physical, role and social functioning and increased fatigue pain and dyspnoea. These findings add to the current evidence available in the literature. Studies of patients with prostate cancer reported significant differences in QLQ-C30 scores, as a result of radiotherapy treatment. Results from one study (Rodrigues, Bezjak et al. 2004) indicated patients experienced significant worsening of fatigue, pain and a deterioration in global health, and another study indicated a significant deterioration in role functioning and increased fatigue (Janda, Gerstner et al. 2000). Kaasa et al (Kaasa, Bjordal et al. 1995) in a study with a heterogenous sample of patients with advanced disease treated with palliative radiotherapy also found significant differences for the physical and role functioning scale and the fatigue scale indicating a deterioration in physical and role functioning and increase in fatigue in this sample. Studies evaluating the QLQ-C30, in which treatment modalities were combined (ie radiotherapy, surgery, hormone therapy, chemotherapy, or combinations of these), also found differences. Aaronson et al (Aaronson, Ahmedzai et al. 1993), in a heterogenous sample of cancer patients, found that for patients whose performance status (as measured by ECOG scores) deteriorated, physical and role functioning and global health deteriorated significantly, and symptoms of fatigue, and nausea and vomiting increased as a result of radiotherapy. Bjordal et al (Bjordal, de Graeff et al. 2000) in their head and neck cancer sample found that all scales except for the emotional functioning scales were significantly different with patients' functioning deteriorating and symptoms worsening, and Krahn et al (Krahn, Bremner et al. 2007) in a prostate cancer sample found small to moderate effect sizes for the physical, role, emotional and social functioning scales showing a deterioration in functioning on these scales.

The instruments were also responsive to the effects of disease recurrence. From recruitment to the first assessment following a diagnosis of disease recurrence, with the exception of the emotional functioning scale, all mean differences were large (15.8 to 31.5) and effect sizes were moderate to large (0.53 to 1.4). By the last assessment following recurrences mean differences had continued to grow and all effects sizes except for the emotional functioning scale were large. For the emotional functioning scale the instruments did pick up differences but these were much smaller (mean difference and effect size of 5.5 and 0.19 respectively) by the last observation following recurrence. Our findings add to those of Bjordal et al (Bjordal, de Graeff et al. 2000)

who, in their study of head and neck cancer patients, also found statistically and clinically significant differences in scores on the QLQ-C30 role, emotional, and social functioning and fatigue and pain scales between patients who had been newly diagnosed and those with a diagnosis of recurrent disease. The QLQ-LC13 dyspnoea scale was also found to be responsive to the effects of disease recurrence with a mean difference and effect size of 20.4 and 0.72 at first observation post recurrence and 31.l6 and 1.11 at last observation.

A limitation of our study was that we focussed predominately on the multi item scales of the instruments, with limited analysis of the single items. This restricted our findings and comparisons to other validation studies. However, we did conduct an extensive array of analyses on the scales to determine validity reliability and responsiveness and as such are confident that our results are valid.

Conclusion

The results of this study confirm that the EORTC QLQ-C30 and the EORTC QLQ-LC13 are valid, reliable and responsive measures of HRQOL in Australians with early stage non-small cell lung cancer. They also add to the already large body of evidence supporting the validity, reliability and responsiveness of the QLQ-C30 as an effective HRQOL measure for a range of cancers, and add to the growing evidence for the QLQ-LC13. In addition, by utilizing data from an existing RCT study in which to conduct this validation study, we have provided an illustration of "validation by application". This is a way to not only add value to HRQOL data collected for other purposes, but perhaps more importantly, to contribute to collective knowledge about the measurement properties and interpretability of HRQOL instruments, particularly in Australian populations and settings.

Appendix

Table A1: Cronbach's α values for the EORTC QLQ-C30 and the EORTC QLQ-LC13 in comparison with previously published data.

Study	Sample	QLQ-C30									QLQ-LC13
		PF	**RF**	**CF**	**EF**	**SF**	**Global**	**Fatigue**	**NV**	**Pain**	**Dyspnoea**
Aaronson et al 1993	Heterogenous	0.68	0.54	0.56	0.73	0.68	0.86	0.80	0.65	0.82	
Ringdal & Ringdal 1993	Heterogenous	0.75	0.55	0.65	0.85	0.72	0.83	-	0.84	0.86	
Bergman et al 1994	Lung										0.81
Fossa et al 1994	Breast	0.61	0.53	0.69	0.75	0.72	0.91	0.83	0.84	0.85	
Osoba et al 1994	Heterogenous	0.68	0.54	0.56	0.73	0.68	0.86	0.80	0.65	0.82	
Kaasa et al 1995	Heterogenous	0.77	0.68	0.62	0.80	0.78	0.88	0.87	0.81	0.89	
Osoba et al 1997	Heterogenous	-	0.69	-	-	-	0.83	-	-	-	
Ringdal et al 1999		0.77	0.63	0.64	0.83	0.76	0.89	0.88	0.77	0.89	
Bjordal et al 2000	Head and neck	0.84	-	-	-	-	-	-	-	-	
Martin et al 2003	Heterogenous	0.78	0.59	0.68	0.85	0.78	0.85	0.77	0.62	0.80	
Chie et al 2004*	Lung										0.83
Galalae et al 2004	Prostate	0.83	0.81	0.61	0.84	0.82	0.87	-	0.89	0.85	
Fitzsimmons et al 2005	Pancreatic	0.83	0.90	0.68	0.87	0.72	0.89	0.82	0.81	0.89	
Galalae et al 2005	Breast	0.60	0.78	0.65	0.81	0.78	0.88	0.82	0.66	0.91	
Luo et al 2005	Heterogenous	0.62	0.87	0.19	0.86	0.83	0.91	0.82	0.68	0.84	
Boehmer et al 2006	Heterogenous	0.78	0.74	0.42	0.84	0.68	0.77	0.86	0.71	0.78	
Poveda et al 2005	Soft tissue sarcoma	0.82	0.74	0.87	0.80	0.80	0.96	0.89	0.97	0.79	
Brabo et al 2006*	Lung										0.79
Nowak et al 2006	Lung	0.52	>0.70	0.51	>0.70	>0.70	>0.70	>0.70	>0.70	>0.70	
Easson et al 2007	Heterogenous	0.87	0.82	0.65	0.84	0.79	0.89	0.87	-	-	
Nicklasson et al 2007	Lung	0.79	0.87	0.57	0.84	0.80	0.87	0.82	0.84	-	0.76
Current study	**Lung**	**0.87**	**0.94**	**0.68**	**0.90**	**0.86**	**0.94**	**0.91`**	**0.67**	**0.87**	**0.89**

Note: *Validation study of a translated version.

References

Aaronson, N. K., S. Ahmedzai, et al. (1993). "The European Organisation for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology." Journal of the National Cancer Institute **85**(5): 365-376.

Aaronson, N. K., S. Ahmedzai, et al. (1991a). The EORTC core quality of life questionnaire: interim results of an international field study. Effect of Cancer on Quality of Life. D. Osoba. Boca Raton, CRC**:** 185-203.

Aaronson, N. K., M. Bullinger, et al. (1988). "A modular approach to quality-of-life assessment in cancer clinical trials." Recent Results in Cancer Research **111**: 231-249.

Aaronson, N. K., A. Cull, et al. (1996). The European Organisation for Research and Treatment of Cancer (EORTC) Modular Approach to Quality of Life Assessment in Oncology: An Update. Quality of Life and Pharmacoeconomics in Clinical Trials. B. Spilker. Philadelphia, Lippincott-Raven Publishers**:** 179-189.

Arbuckle, J. and W. Wothke (1999). Amos 4.0 users guide. Chicago, SPSS/Small Waters.

Bentler, P. (1990). "Comparitive Fit Index in structural models" Psychological Bulletin **107**(2): 238-249.

Bergman, B., N. K. Aaronson, et al. (1994). "The EORTC QLQ-LC13: a modular supplement to the EORTC core quality of life questionnaire (QLQ-C30) for use in lung cancer clinical trials." European Journal of Cancer **30A**(5): 635-642.

Berlangieri, S. U. and A. M. Scott (2000). "Metabolic staging of lung cancer." New England Journal of Medicine **343**(4): 290-292.

Bjordal, K., A. de Graeff, et al. (2000). "A 12 country field study of the EORTC QLQ-C30 (version 3.0) and the head and neck cancer specific module (EORTC QLQ-H&N35) in head and neck patients. EORTC Quality of Life Group." European Journal of Cancer **36**(14): 1796-807.

Blazeby, J. M., T. Conroy, et al. (2003). "Clinical and psychometric validation of an EORTC questionnaire module, the EORTC QLQ-OES18, to assess quality of life in patients with oesophageal cancer." European Journal of Cancer **39**(10): 1384-94.

Boehmer, S. and A. Luszczynska (2006). "Two kinds of items in quality of life instruments: 'indicator and causal variables' in the EORTC qlq-c30." Quality of Life Research **15**(1): 131-41.

Brabo, E. P., M. E. Paschoal, et al. (2006). "Brazilian version of the QLQ-LC13 lung cancer module of the European Organization for Research and Treatment of Cancer: preliminary reliability and validity report." Quality of Life Research **15**(9): 1519-24.

Browne, M. and R. Cudeck (1993). Alternative ways of assessing model fit. Testing structural equation models. K. Bollen and J. Long. Newbury Park, CA, Sage Publications Inc**:** 136-162.

Cella, D. F. and D. S. Tulsky (1993b). "Quality of life in cancer: definition, purpose, and method of measurement." Cancer Investigation **11**(3): 327-36.

Chie, W.-C., C.-H. Yang, et al. (2004). "Quality of life of lung cancer patients: validation of the Taiwan Chinese version of the EORTC QLQ-C30 and QLQ-LC13." Quality of Life Research **13**: 257-262.

Chie, W. C., K. J. Chang, et al. (2003). "Quality of life of breast cancer patients in Taiwan: validation of the Taiwan Chinese version of the EORTC QLQ-C30 and EORTC QLQ-BR23." Psycho-Oncology **12**(7): 729-35.

Cleton, F. J. (1995). Chemotherapy: general aspects. Oxford Textbook of Oncology. M. Peckham, H. Pinedo and U. Veronesi. Oxford, Oxford University Press. **2:** 445-452.

Cocks, K., D. Cohen, et al. (2007). "An international field study of the reliability and validity of a disease-specific questionnaire module (the QLQ-MY20) in assessing the quality of life of patients with multiple myeloma." European Journal of Cancer **43**(11): 1670-8.
Cohen, J. (1988). Statistical Power Analysis for the Behavioural Sciences. Hillsdale, NJ, Lawrence Erlbaum Associates.
Deyo, R. A., P. Diehr, et al. (1991). "Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation." Controlled Clinical Trials **12**(4 Suppl): 142S-158S.
Dunn, G. (1992). "Design and analysis of reliability studies." Statistical Methods in Medical Research **1**(2): 123-57.
Easson, A. M., A. Bezjak, et al. (2007). "The ability of existing questionnaires to measure symptom change after paracentesis for symptomatic ascites." Annals of Surgical Oncology **14**(8): 2348-57.
Fallowfield, L. (1990). The quality of life in cancer. The Quality of Life: The Missing Measurement in Health Care. L. Fallowfield. London, Souvenir**:** 75-113.
Fayers, P., N. Aaronson, et al. (2001). The EORTC QLQ-C30 Scoring Manual. Brussels, European Organisation for Research and Treatment of Cancer.
Fayers, P. M., N. K. Aaronson, et al. (1999). EORTC QLQ-C30 Scoring Manual. Brussels, EORTC Study Group on Quality of Life.
Fayers, P. M. and D. Machin (2000). Quality of Life: Assessment, Analysis and Interpretation, John Wiley & Sons Ltd.
Fayers, P. M., S. Weeden, et al. (1998). EORTC QLQ-C30 Reference Values. Brussels, EORTC Study Group on Quality of Life.
Fitzsimmons, D., S. Kahl, et al. (2005). "Symptoms and quality of life in chronic pancreatitis assessed by structured interview and the EORTC QLQ-C30 and QLQ-PAN26." American Journal of Gastroenterology **100**(4): 918-26.
Fossa, S. D. (1994). "Quality of life assessment in unselected oncologic out-patients: A pilot study." International Journal of Oncology **4**: 1393-1397.
Galalae, R. M., T. Loch, et al. (2004). "Health-related quality of life measurement in long-term survivors and outcome following radical radiotherapy for localized prostate cancer." Strahlentherapie und Onkologie **180**(9): 582-9.
Galalae, R. M., J. Michel, et al. (2005). "Significant negative impact of adjuvant chemotherapy on health-related quality of life (HR-QoL) in women with breast cancer treated by conserving surgery and postoperative 3-D radiotherapy. A prospective measurement." Strahlentherapie und Onkologie **181**(10): 645-51.
Greco, M. (1995). Achievements and obstacles to progress in cancer surgery. Oxford Textbook of Oncology. M. Peckham, H. Pinedo and U. Veronesi. Oxford, Oxford University Press. **1:** 865-879.
Guyatt, G. H., R. Jaeschke, et al. (1996). Measurements in clinical trials: Choosing the right approach. Quality of Life and Pharmacoeconomics in Clinical Trials. R. Spilker. New York, Lippincott-Raven**:** 41-48.
Hanks, G. W. and P. J. Hoskin (1995). Pain and symptom control in advanced cancer. Oxford Textbook of Oncology. M. Peckham, H. Pinedo and U. Veronesi. Oxford, Oxford University Press. **2:** 2417-2430.
Hatcher, L. (1994). A Step-by-Step Approach to Using the SAS System for Factor Analysis and Structural Equation Modeling. Cary, N.C. , SAS Institute Inc.
Hattie, J. A. (1985). "Methodology review: Assessing unidimensionality of tests and items." Applied Psychological Measurement **9**(2): 139-164.

Hu, L.-T. and P. Bentler (1999). "Cut-off criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives." Structural equation modelling **6**: 1-55.
Janda, M., N. Gerstner, et al. (2000). "Quality of life changes during conformal radiation therapy for prostate carcinoma." Cancer **89**(6): 1322-8.
Joreskog, K. (1993). Testing structural equation models. Testing structural equation models. K. Bollen and J. Long. Newbury park, CA, Sage Publications Inc**:** 294-316.
Kaasa, S., K. Bjordal, et al. (1995). "The EORTC core quality of life questionnaire (QLQ-C30): validity and reliability when analysed with patients treated with palliative radiotherapy." European Journal of Cancer **31A**(13-14): 2260-3.
Kazis, L. E., J. J. Anderson, et al. (1989). "Effect sizes for interpreting changes in health status." Medical Care **27**(3 Suppl): S178-89.
Kenny, P., M. King, et al. (2008). "Quality of life and survival in the two years after surgery for non-small cell lung cancer." Journal of Clinical Oncology **26**(2): 233-241.
King, M. T. (1996). "The interpretation of scores from the EORTC quality of life questionnaire QLQ-C30." Quality of Life Research **5**(6): 555-67.
Krabbe, P. F., L. Peerenboom, et al. (2004). "Responsiveness of the generic EQ-5D summary measure compared to the disease-specific EORTC QLQ C-30." Quality of Life Research **13**(7): 1247-53.
Krahn, M., K. E. Bremner, et al. (2007). "Responsiveness of disease-specific and generic utility instruments in prostate cancer patients." Quality of Life Research **16**(3): 509-22.
Liang, M. H., A. H. Fossel, et al. (1990). "Comparisons of five health status instruments for orthopedic evaluation." Medical Care **28**(7): 632-42.
Liang, M. H., M. G. Larson, et al. (1985). "Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research." Arthritis & Rheumatism **28**(5): 542-7.
Lohr, K. N., N. K. Aaronson, et al. (1996). "Evaluating quality-of-life and health status instruments: development of scientific review criteria." Clinical Therapeutics **18**(5): 979-992.
Lundh Hagelin, C., A. Seiger, et al. (2006). "Quality of life in terminal care-with special reference to age, gender and marital status." Supportive Care in Cancer **14**(4): 320-8.
Luo, N., C. S. Fones, et al. (2005). "The European Organization for Research and Treatment of Cancer Quality of Life Questionnaire (EORTC QLQ-c30): validation of English version in Singapore." Quality of Life Research **14**(4): 1181-6.
Maguire, P. (1995). Psychological sequelae of cancer and its treatment. Oxford Textbook of Oncology. M. Peckham, H. Pinedo and U. Veronesi. Oxford, Oxford University Press. **2:** 2408-2416.
Martin, C. G., E. B. Rubenstein, et al. (2003). "Measuring chemotherapy-induced nausea and emesis." Cancer **98**(3): 645-55.
McLachlan, S. A., G. M. Devins, et al. (1998). "Validation of the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire (QLQ-C30) as a measure of psychosocial function in breast cancer patients." European Journal of Cancer **34**(4): 510-7.
Michelson, H., C. Bolund, et al. (2000). "Multiple chronic health problems are negatively associated with health related quality of life (HRQoL) irrespective of age." Quality of Life Research **9**(10): 1093-104.
Nicklasson, M. and B. Bergman (2007). "Validity, reliability and clinical relevance of EORTC QLQ-C30 and LC13 in patients with chest malignancies in a palliative setting." Quality of Life Research **16**(6): 1019-28.

Nolop, K. B., C. G. Rhodes, et al. (1987). "Glucose utilization in vivo by human pulmonary neoplasms." Cancer **60**(11): 2682-9.
Nowak, A. K., M. R. Stockler, et al. (2004). "Assessing quality of life during chemotherapy for pleural mesothelioma: feasibility, validity, and results of using the European Organization for Research and Treatment of Cancer Core Quality of Life Questionnaire and Lung Cancer Module." Journal of Clinical Oncology **22**(15): 3172-3180.
Nunnally, J. C. (1978). Psychometric Theory, McGraw-Hill.
Osoba, D. (1995). "Measuring the effect of cancer on health-related quality of life." Pharmacoeconomics **7**(4): 308-19.
Osoba, D., N. K. Aaronson, et al. (1991). A practical guide for selecting quality-of-life measures in clinical trials and practice. Effect of Cancer on Quality of Life. D. Osoba. Boca Raton, CRC**:** 90-103.
Osoba, D., B. Zee, et al. (1994). "Psychometric properties and responsiveness of the EORTC Quality of Life Questionnaire (QLQ-C30) in patients with breast, ovarian and lung cancer." Quality of Life Research **3**(5): 353-64.
Pieterman, R. M., J. W. van Putten, et al. (2000). "Preoperative staging of non-small-cell lung cancer with positron-emission tomography." New England Journal of Medicine **343**(4): 254-61.
Poveda, A., A. Lopez-Pousa, et al. (2005). "Phase II clinical trial with pegylated liposomal doxorubicin (CAELYX/Doxil) and quality of life evaluation (EORTC QLQ-C30) in adult patients with advanced soft tissue sarcomas: A study of the Spanish Group for Research in Sarcomas (GEIS)." Sarcoma **9**(3-4): 127-132.
Ringdal, G. I. and K. Ringdal (1993). "Testing the EORTC quality of life questionnaire on cancer patients with heterogeneous diagnoses." Quality of Life Research **2**: 129-140.
Ringdal, K., G. I. Ringdal, et al. (1999). "Assessing the consistency of psychometric properties of the HRQoL scales within the EORTC QLQ-C30 across populations by means of the Mokken Scaling Model." Quality of Life Research **8**(1-2): 25-43.
Robert, G. and R. Milne (1999). "A Delphi study to establish national cost-effectiveness research priorities for positron emission tomography." European Journal of Radiology **30**: 54-60.
Rodrigues, G., A. Bezjak, et al. (2004). "The relationship of changes in EORTC QLQ-C30 scores to ratings on the Subjective Significance Questionnaire in men with localized prostate cancer." Quality of Life Research **13**(7): 1235-46.
Saunders, C. A., J. E. Dussek, et al. (1999). "Evaluation of fluorine-18-fluorodeoxyglucose whole body positron emission tomography imaging in the staging of lung cancer." Annals of Thoracic Surgery **67**(3): 790-7.
Schipper, H. and J. Clinch (1988). Assessment of treatment in cancer. Measuring Health: a Practical Approach. G. Teeling Smith. Chichester, John Wiley**:** 109-155.
Schipper, H., J. J. Clinch, et al. (1996). Quality of life studies: definitions and conceptual issues. Quality of Life and Pharmacoeconomics in Clinical Trials. B. Spilker. New York, Lippincott-Raven**:** 11-23.
Spilker, B., Ed. (1996). Quality of Life and Pharmacoeconomics in Clinical Trials. Philadelphia, Lippincott-Raven Publishers.
Steel, G. G. (1995). The biological basis of radiotherapy. Oxford Textbook of Oncology. M. Peckham, H. Pinedo and U. Veronesi. Oxford, Oxford University Press. **1:** 668-680.
Stockler, M. R., D. Osoba, et al. (1999). "Convergent discriminitive, and predictive validity of the Prostate Cancer Specific Quality of Life Instrument (PROSQOLI) assessment and comparison with analogous scales from the EORTC QLQ-C30 and a trial-specific module. European Organisation for Research and Treatment of Cancer. Core Quality of Life Questionnaire." Journal of Clinical Epidemiology **52**(7): 653-66.

Streiner, D. L. and G. R. Norman (1996). Health Measurement Scales: a Practical Guide to their Development and Use. Oxford, Oxford University Press.
Tabachnick, B. G. and L. S. Fidell (2001). Using multivariate statistics Boston, Massachusetts, Allyn and Bacon.
Tanaka, J. (1993). Multifaceted conceptions of fit in structural equation models. Testing structural models. K. A. Bollen and J. Long. Newbury Park, CA, Sage Publications**:** 10-39.
Viney, R., M. Boyer, et al. (2004). "Randomized controlled trial of the role of Positron Emission Tomography in the management of stage I and II Non-Small-Cell lung cancer." Journal of Clinical Oncology **22**: 2357-2362.
Wahl, R. L., L. E. Quint, et al. (1994). "Staging of mediastinal non-small cell lung cancer with FDG PET, CT, and fusion images: preliminary prospective evaluation." Radiology **191**(2): 371-7.
Ware, J. E. (1987). "Standards for validating health measures: definition and content." Journal of Chronic Diseases **40**(6): 473-480.
Weder, W., R. A. Schmid, et al. (1998). "Detection of extrathoracic metastases by positron emission tomography in lung cancer." Annals of Thoracic Surgery **66**(3): 886-92; discussion 892-3.
Wood-Dauphinee, S. (1999). "Assessing quality of life in clinical research: from where have we come and where are we going?" Journal of Clinical Epidemiology **52**(4): 355-63.