VALIDATION OF THE PROSQOLI
AS AN OUTCOME MEASURE FOR CLINICAL TRIALS IN
ADVANCED HORMONE-RESISTANT PROSTATE CANCER:
ASSESSMENT OF CONVERGENT, DISCRIMINATIVE AND PREDICTIVE
VALIDITY WITH BASELINE DATA FROM A RANDOMISED TRIAL.

by

Martin Richard Stockler

A thesis submitted in conformity with the requirements
for the degree of Master of Science in Clinical Epidemiology
Graduate Department of Community Health
University of Toronto

© Copyright by Martin Richard Stockler 1996
The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author’s permission.

L’auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L’auteur conserve la propriété du droit d’auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-51552-4
VALIDATION OF THE PROSQOLI AS AN OUTCOME MEASURE FOR CLINICAL TRIALS IN ADVANCED HORMONE-RESISTANT PROSTATE CANCER: ASSESSMENT OF CONVERGENT, DISCRIMINATIVE AND PREDICTIVE VALIDITY WITH BASELINE DATA FROM A RANDOMISED TRIAL.

Martin Richard Stockler. Master of Science in Clinical Epidemiology, 1996.
Graduate Department of Community Health, University of Toronto.

ABSTRACT
The Prostate Cancer Specific Quality of Life Instrument (PROSQOLI) was designed to be an outcome measure for clinical trials in advanced hormone-resistant prostate cancer. A valid measure of health-related quality of life (HRQL) is essential in this setting. This thesis describes the validation of the PROSQOLI using baseline data from a randomised trial in which HRQL was also assessed with the European Organisation for Research and Treatment of Cancer (EORTC) Core Quality of Life Questionnaire. Convergent validity assessed with the multitrait-multimethod matrix approach, discriminative validity assessed according to conventional clinical criteria, and the ability to predict survival duration provided strong support for the validity of all PROSQOLI scales except those for family/marriage relationships and passing urine. The strength, consistency and independence of the prognostic information provided by the HRQL scales was striking. This thesis provides strong support for the cross-sectional validity of the scales assessing pain, fatigue, appetite, constipation, physical activity, mood, and overall well-being and indicates the need to modify the scales for family/marriage relationships and passing urine.
ACKNOWLEDGEMENTS

Ian Tannock for the opportunity and encouragement to do this work.

Pamela Goodwin for ensuring that I finished.

Paul Corey for making me think about what I was doing.

Claire Bombardier for making me think about why I was doing it.

David Osoba for helpful comments.

David Bloomfield, Andreas Maetzel & Sue-Anne McLachlan for limitless discussions.

Jaine and Isobel Stockler for inspiration, patience and support.

The men who participated in the randomized trial.
TABLE OF CONTENTS

ABSTRACT ................................................................. i

ACKNOWLEDGEMENTS .................................................. ii

TABLE OF CONTENTS ................................................... iii

1 BACKGROUND .......................................................... 1

  1.1 Advanced prostate cancer and its treatment ..................... 1
  1.2 Assessment of quality of life .................................... 2
     1.2.1 Introduction .............................................. 2
     1.2.2 The generic-specific spectrum ........................... 2
     1.2.3 Quality-of-life assessment in cancer research .......... 3
     1.2.4 Quality-of-life assessment in prostate cancer ......... 4
  1.3 Objectives of the current study ................................ 10
  1.4 The randomised trial ............................................ 10
  1.5 The PROSQOLI .................................................. 13
     1.5.1 Development and prior validation of the PROSQOLI .... 13
     1.5.2 Description of the PROSQOLI (see appendix 1) ......... 14
  1.6 The EORTC instrument (EORTCI) ................................ 16
     1.6.1 Development and prior validation of the EORTC QLQ-C30 .. 16
     1.6.2 The QLM-P14 .............................................. 19
     1.6.3 Description of the EORTCI (see appendices 2 and 3) .... 19
3 DESCRIPTIVE DATA ................................................................. 32

3.1 Introduction and aims ......................................................... 32

3.2 Methods ............................................................................. 32

3.2.1 Preparation of the data .................................................... 32

3.2.2 General descriptive methods ............................................ 33

3.2.3 Compliance with the patient-based measures ...................... 33

3.2.4 Transformation of the patient-based measures ................. 33

3.2.5 Non-parametric statistics ............................................... 34

3.3 Findings .............................................................................. 36

3.3.1 Description of patients .................................................... 36

3.3.2 Conventional measures .................................................... 36

3.3.2.1 Compliance with conventional measures ....................... 36

3.3.2.2 Distributions of conventional measures ......................... 36

3.3.3 Patient-based measures ..................................................... 38

3.3.3.1 Compliance with patient-based measures ....................... 38

3.3.3.2 Distributions of the patient-based measures ................... 41

3.4 Discussion .......................................................................... 50

3.5 Conclusions ......................................................................... 51
# 4 CONVERGENT VALIDITY

## 4.1 Introduction and Aims

## 4.2 Methods

### 4.2.1 Bivariate relationships

#### 4.2.1.1 The multitrait-multimethod matrix approach

#### 4.2.1.2 Index of association

#### 4.2.1.3 Interpretation of the correlation coefficients

#### 4.2.1.4 Hypotheses

### 4.2.2 Multivariable relationships

#### 4.2.2.1 Introduction and aims

#### 4.2.2.2 Statistical methods

## 4.3 Findings

### 4.3.1 Bivariate relationships

#### 4.3.1.1 Description of the correlation matrix

#### 4.3.1.2 Validity diagonal

#### 4.3.1.3 Specific scales and global scales

#### 4.3.1.4 Measures of pain

#### 4.3.1.5 Related but distinct domains

### 4.3.2 Multivariable relationships

## 4.4 Discussion

## 4.5 Conclusions
6 PREDICTIVE VALIDITY ................................................................. 93
6.1 Introduction ............................................................................ 93
6.2 Univariable Methods ............................................................... 93
  6.2.1 Introduction and aim ......................................................... 93
  6.2.2 Hypotheses ....................................................................... 93
  6.2.3 Dichotomisation of predictor variables .............................. 94
  6.2.4 Comparison of survival distributions ............................... 98
  6.2.5 Expression of the results ................................................... 98
6.3 Multivariable Methods ............................................................ 100
  6.3.1 Introduction and aims ....................................................... 100
  6.3.2 Adjusted "univariable" analyses ......................................... 100
  6.3.3 Patient-based measures alone .......................................... 100
  6.3.4 Patient-based and conventional measures ....................... 100
  6.3.5 Subset selection ............................................................... 101
  6.3.6 Continuous predictor variables ......................................... 101
  6.3.7 Regression model ............................................................ 102
  6.3.8 Expression of the results ................................................... 105
6.4 Findings .................................................................................. 107
  6.4.1 Simple univariable associations with survival ................... 107
  6.4.2 Adjusted "univariable" analyses ......................................... 121
  6.4.3 Multivariable prognostic significance .............................. 122
6.5 Discussion ................................................................................ 124
6.6 Conclusions ............................................................................ 124
1 BACKGROUND

1.1 Advanced prostate cancer and its treatment

Advanced prostate cancer is common, incurable and causes substantial suffering. It is the most common cancer and the 2nd commonest cause of cancer death in Canadian men: the estimated age-standardised incidence and mortality rates in Canada are 65 and 18 per 100,000 men per year. More than 50% of men with prostate cancer develop advanced disease which is characterised by either local extension, distant spread or both, and is incurable. The majority of men - about 80% - with advanced prostate cancer initially respond to androgen withdrawal with improvement in pain and other symptoms. However these responses are transient lasting an average of one year before hormone-resistance develops. The median survival of men with advanced hormone-resistant disease is about 1 year.

Progress in the treatment of advanced hormone-resistant prostate cancer has been hampered by a lack of suitable outcome measures. Clinical trials in patients with advanced cancer have focused most often on indices of tumour shrinkage. Advanced prostate cancer is characterised by sclerotic bony metastases and locoregional pelvic disease in which changes over time are difficult to assess. The serum concentration of Prostate Specific Antigen (PSA) appears to be associated with tumour bulk, and change in serum PSA has been proposed as an outcome measure for the assessment of response to hormones or chemotherapy. However indices of change in tumour bulk are, at best, only surrogate measures of patient benefit. The goals of treatment in advanced prostate cancer are palliative - to relieve symptoms and prolong life without prospect of cure. Therefore the criteria by which treatments are judged should be quality and quantity of
Duration of survival is assessed in most clinical trials; however, available treatments for advanced hormone-resistant prostate cancer are unlikely to have a major impact on survival. Treatments for advanced hormone-resistant prostate cancer may have a profound effect on quality-of-life, whether by relieving cancer-related symptoms or by adding those of treatment toxicity. Therefore, suitable measures of the effects of disease and treatment on quality of life are essential for the evaluation of treatments in this setting.

1.2 Assessment of quality of life

1.2.1 Introduction

Quality-of-life is an abstract, multi-dimensional construct reflecting physical, psychological and social aspects, which includes, but is not limited to, the concept of health.\textsuperscript{8,9,10,11} It reflects an individual's perception of, and response to, their unique circumstances.\textsuperscript{12} This definition gives primacy to the individual's views and identifies self-assessment as optimal. Separate, parallel traditions of quality of life research have developed in health services research and clinical research. Differences between these traditions reflect differences in the backgrounds and goals of those in each field.

1.2.2 The generic-specific spectrum

The most generic quality of life instruments, developed for demographically-defined and therefore heterogenous populations, are largely the product of health services researchers with backgrounds in sociology and psychology whose primary interests are quality of life and public health (see for example the Medical Outcomes Study Short Form 36\textsuperscript{13} and the Sickness Impact Profile\textsuperscript{14}). More specific instruments, developed for clinically-defined and therefore more homogeneous populations, are largely the product of clinical researchers with biomedical backgrounds whose primary interests are the management of
specific diseases. Generic instruments should be applicable in a wide range of applications and therefore provide the opportunity for comparisons between groups with quite different characteristics. Specific instruments should be focussed on the problems of the population for which they were developed and therefore should provide a more pertinent description than generic instruments. Whether these expectations hold true will depend on the particular instruments and research questions. Although it is easier to argue in terms of dichotomies, quality of life instruments are probably better described by their position on a spectrum from highly generic to highly specific; where a particular instrument lies on this spectrum is a matter of perspective. The same instrument may be too specific for one application but not specific enough for another. An instrument designed to assess quality of life in men with advanced prostate cancer may be too specific for a study looking at all men with prostate cancer and too general for a study looking at men with end-stage disease. Arguments based on simple generalities may be insufficient to answer complex specific questions: what is best in a particular study depends on what the investigators want to know.

1.2.3 Quality-of-life assessment in cancer research

Many quality of life instruments have been developed by and for cancer researchers interested in overlapping but different populations of people who have been affected by cancer. Instruments can be described in terms of the target population for which they were developed. Examples from the most generic through increasing degrees of specificity include: the Cancer Rehabilitation and Evaluation Systems Inventory (CARES) - developed for people who have had cancer; the Functional Living Index Cancer (FLIC) - developed for people with cancer; the European Organisation for Research and
Treatment of Cancer Core Quality of Life Questionnaire (EORTC QLQ-C30) developed for people with cancer participating in international clinical trials;\textsuperscript{17} the Functional Assessment of Cancer Therapy - General (FACT-G) - developed for people receiving cancer treatments in general;\textsuperscript{18} the GLQ-8 - developed for people receiving chemotherapy for cancer;\textsuperscript{19} and the Breast Cancer Chemotherapy Questionnaire - developed for women having adjuvant chemotherapy for breast cancer.\textsuperscript{20} Acknowledging the tension between generalizability and specificity, several groups have recommended a modular approach incorporating instruments which reconcile these conflicting needs by adding specific items to a generic core.\textsuperscript{17,18,21}

1.2.4 Quality-of-life assessment in prostate cancer

Prostate cancer is a heterogeneous disease, in both its biological behaviour and its effect on the host. The aspects of quality of life which are most important to men with asymptomatic early stage disease considering local treatments are likely to be different to those of men with more advanced disease considering androgen withdrawal therapy which are likely to be different again to those of men with symptomatic hormone-resistant disease. So far, most of the questionnaires used in patients with advanced disease have been developed by clinical researchers interested in specific questions, such as differences between various methods of androgen withdrawal for newly diagnosed advanced disease or the effects of palliative chemotherapy for advanced hormone-resistant disease. As a consequence, most of the instruments are not only disease-specific, but also stage- and treatment-specific resulting in considerable heterogeneity among the items covered by different instruments.
Much of the published literature about quality of life in men with advanced prostate cancer comes from studies which used questionnaires based on the work of the EORTC Study Group on Quality of Life and the EORTC Genitourinary Group. These studies were performed during the development of the EORTC core quality of life questionnaire (QLQ-C30). The instruments used in these studies include varying combinations of scales that were later included in the QLQ-C30 (such as physical function, emotional function and social function) as well as scales that were created specifically for men with prostate cancer (such as urinary symptoms and symptoms of androgen withdrawal). These studies are largely descriptive and most of them address issues of validity only indirectly.

In 1989 Fossa et al reported results with a 33-item EORTC questionnaire assessing physical function, psychological distress, social functioning, role functioning, sexual function, urinary symptoms, pain, fatigue, nausea, vomiting, diarrhoea and insomnia (but no global items) which was administered to 67 men with hormone resistant prostate cancer admitted to hospital for palliative radiation therapy. Pain, fatigue, sexual dysfunction and interference with social function were the most frequently reported problems. It was claimed that the validity of these scales was supported by correlations with independent measures such as physician-rated performance status, hemoglobin concentration, analgesic use, and the degree of bony involvement but the details were not shown. This study also provided evidence of the discordance between physician and patient assessments of physical function, with the physicians consistently underestimating how badly their patients were affected by their disease.

The following year Fossa et al reported results with the same questionnaire administered to 72 of 171 men with advanced hormone resistant prostate cancer
participating in a randomised trial of mitomycin-C versus estramustine in which the HRQL assessment was optional. Decreased functional status, fatigue and pain were the most highly rated problems and rarely were relieved by treatment. Significant urinary symptoms were not reported by any of the 72 patients completing the questionnaire.

The most recent report from Fossa’s group describes the results with a shorter 19-item questionnaire restricted to physical and role function, pain, fatigue, nausea, vomiting, diarrhoea, and global perceptions which was administered to 137 men with advanced hormone resistant prostate cancer admitted for radiation therapy or strontium-89 to palliate painful bony metastases. Item-scale and inter-scale correlations supported the validity of the questionnaire. Pain, physical function, and performance status were significant independent predictors of survival duration. The authors were disappointed by the lack of improvement following palliative radiation therapy or strontium; however, there was no control group which did not receive these treatments for comparison.

In summary, these reports from Fossa and her colleagues describe quality of life in 3 separate series of men with advanced hormone-resistant prostate cancer using different versions of the EORTC instrument. There was poor agreement between physicians’ and patients’ assessments of sexual function and physical status. Urinary symptoms were not prominent in the two studies in which they were assessed although the same questionnaire detected urinary symptoms in men with bladder cancer.

Moore et al reported improvements in emotional function, social function, pain, and anorexia as assessed by the EORTC QLQ-C30 and independent measures of pain in 27 men with advanced hormone resistant prostate cancer treated with mitoxantrone and low dose prednisone. The cross-sectional validity of the instruments was not addressed.
Tannock et al described results using a combination of 17 linear analog scales, the McGill Pain Questionnaire and an index of analgesic use, in 37 men with advanced hormone resistant prostate cancer. Pain, fatigue, reduced overall well-being, depressed mood and anxiety were the most prominent symptoms. Pain improved in 38% of the men and improvements in other dimensions of quality of life were seen in those men who had improvements in pain. Inter-scale correlations supported the validity of the instrument and were used as a basis for item reduction. This instrument was the precursor of the Prostate Cancer Specific Quality of Life Instrument (PROSQOLI) the validity of which is the primary focus of this thesis.

A larger literature exists on the quality of life of men with newly diagnosed advanced prostate cancer. The major focus of these studies has been on the effects of androgen withdrawal.

Herr, Kornblith et al reported preliminary results with a 49 item questionnaire based on the EORTC instrument in a group of 35 men with previously untreated metastatic disease who were considering androgen withdrawal therapy. In a subsequent study these researchers assessed the quality of life of 172 men with the whole spectrum of prostate cancer by administering questionnaires to the patients and 83 of their wives. Of the 110 men with complete data, 61 were receiving hormonal therapy, 14 were receiving no treatment, and 35 had received only local therapy (radiation therapy or prostatectomy). Sexual dysfunction, fatigue, urinary problems and pain were the most prominent problems. Those receiving androgen withdrawal therapy reported much poorer sexual function. Interestingly, urinary symptoms (in the patient) seemed to be more important to the wives than to the men themselves: the wives' ratings of their husbands' urinary symptoms and
global quality of life were correlated more highly than their husbands' own ratings of urinary symptoms and global quality of life.

Cleary et al described the baseline quality of life scores in 487 men with previously untreated metastatic prostate cancer who participated in 2 trials comparing different methods of androgen withdrawal. The 34-item questionnaire consisted of scales from a number of instruments including the RAND Medical Outcomes Study Short Form General Health Survey (MOS-SF20), the CARES inventory and the US National Health Interview Survey instrument. Vitality, social functioning, emotional functioning and physical capacity were the domains most highly correlated with global perceptions. There was little correlation between sexual interest or sexual functioning and global perceptions. There were no questions regarding urinary symptoms.

Cassileth et al assessed HRQL in 147 men with metastatic prostate cancer who had been treated with androgen withdrawal. Androgen withdrawal can be achieved by surgical removal of the testes (surgical castration) or by drug therapy (medical castration). Medical castration had been used in 115 of the men and surgical castration in the other 32 men. HRQL was assessed with the FLIC, the Profile of Mood States and additional items addressing physical, interpersonal and sexual issues. Androgen withdrawal was associated with a marked deterioration in sexual activity and interest, and modest improvements in psychosocial status and global quality of life.

The above studies attest to the feasibility of HRQL assessment in men with prostate cancer. There seems to be consensus that physical function, emotional function, social function, pain and fatigue should be assessed. Sexual function and urinary symptoms were measured more often, scored higher, and were more highly correlated with global
perceptions in men with earlier stage disease, whereas pain, fatigue, anorexia and functional impairment were more prominent in men with later stage disease. Constipation was not assessed often, although it may be important in men with advanced disease taking narcotic analgesics.
1.3 Objectives of the current study

In order to achieve its intended status as a scientific instrument, the PROSQOLI must be shown to be valid. In other words: does it measure what it is supposed to? This thesis describes the results of a formal assessment of the empirical validity of the PROSQOLI using baseline data from a randomized trial of prednisone versus prednisone plus mitoxantrone for symptomatic men with advanced, hormone-resistant prostate cancer. In this trial quality of life was assessed with both the PROSQOLI and a modular instrument based on the EORTC Core Quality of Life Questionaire (EORTC QLQ-C30). This thesis addresses the distributional, convergent, discriminative and predictive validity of the PROSQOLI in comparison with the analogous scales from the EORTC instrument.

The distributions of responses are examined using standard graphical and summary techniques in order to learn about the instruments and the study population. Convergent validity is assessed with a multitrait-multimethod matrix of the rank correlations between the scales of the PROSQOLI and analogous scales from the EORTC instrument. The relationships between the specific scales and the global ratings is further explored with multiple regression. Discriminative validity is assessed by evaluating the ability of the PROSQOLI to detect anticipated differences in HRQL in groups defined according to independent clinical criteria. Predictive validity is assessed by evaluating the ability of the PROSQOLI scales to predict survival duration.

1.4 The randomised trial

Full details of the randomized trial on which this analysis is based are reported elsewhere. In brief, 161 symptomatic men with advanced hormone-resistant prostate cancer were recruited from 10 Canadian institutions and allocated randomly to receive
treatment with either daily low-dose oral prednisone plus mitoxantrone chemotherapy or daily low-dose oral prednisone alone.

Patients were assessed every 3 weeks with 1) the PROSQOLI which includes the 6 point ordinal adjectival scale of Present Pain Intensity from the McGill Pain Questionnaire (the PPI scale), 9 linear analog self assessment scales (LASAS) and an index of daily analgesic intake (the Analgesic Score), and 2) the EORTC instruments (EORTC1) which included the 30-item core quality of life questionnaire (QLQ-C30) and a 14 item trial-specific module (QLM-P14).

The primary endpoint for the trial was a sustained and substantial improvement in pain intensity without any increase in analgesic intake. In detail, the primary endpoint was defined as a $\geq 2$ point improvement in the 6 point ordinal adjectival scale of Present Pain Intensity (PPI) or complete disappearance of pain, maintained for at least 2 consecutive assessments at least 3 weeks apart without any increase in the Analgesic Score. Any man randomised who did not satisfy the primary endpoint was classified as a non-responder.

Twenty-three of the 80 men in the prednisone plus mitoxantrone group and 10 of 81 men in the prednisone alone group satisfied the primary endpoint (29% versus 12%, $p = 0.01$). The alternative measures of pain from the PROSQOLI and EORTC1 corroborated the finding of better palliation of pain in those men treated with prednisone plus mitoxantrone. The duration of palliative response was substantially longer for men treated with prednisone plus mitoxantrone than prednisone alone (11 versus 4 months, $p < .0001$). There were substantial improvements in most domains of quality of life in patients who satisfied the primary endpoint.

The following sections describe the two patient based measures - the PROSQOLI
and the EORTCI and then the conventional clinical measures used in the randomised trial and considered in this thesis.
1.5 The PROSQOLI

1.5.1 Development and prior validation of the PROSQOLI

The Prostate Cancer Specific Quality-of-life Instrument (PROSQOLI) was developed to be a pragmatic outcome measure for clinical trials of systemic treatments in symptomatic men with advanced hormone-resistant prostate cancer. The PROSQOLI is designed to assess the palliative benefit of systemic treatments in terms of their effects on health-related quality-of-life. It is therefore a measure of disease-specific, health-related quality-of-life (HRQL) and as such, it is deliberately focused on symptoms, function and global perceptions specific to advanced hormone-resistant prostate cancer and its treatment. It was designed to be brief, simple and suitable for repeated administration to a population of sick, elderly men on narcotic analgesics, for many of whom English was a second language. It consists of 9 linear analogue self-assessment scales (LASAS) assessing symptoms, function, and overall well-being; a 6-point ordinal adjectival scale of pain intensity based on the Present Pain Intensity scale from the McGill Pain Questionnaire\textsuperscript{36,37} (the PPI) and a summative index of analgesic intake (the Analgesic Score).

The PROSQOLI was developed and piloted in two single-arm studies. The first study assessed the effects of low-dose oral prednisone in 37 men with symptomatic advanced HRPC.\textsuperscript{26} Pain, the dominant symptom, was assessed with the long version of the McGill Pain Questionnaire with a single LASAS and with the Analgesic Score. Other aspects of HRQL were assessed with 16 LASAS adapted from an instrument developed for women having chemotherapy for breast cancer.\textsuperscript{38,39} Administration of the 6-point adjectival Present Pain Intensity (PPI) scale from the McGill Pain Questionnaire and the pain LASAS was feasible and yielded consistent results. The more complex Pain Rating
Index of the McGill Pain Questionnaire required considerable language skills and was difficult for patients whose first language was not English. Many of the LASAS were highly correlated (for example mood with anxiety and fatigue with sleeping), whilst others addressed aspects rarely rated as troublesome (for example mental function, housework and speaking/writing). On the basis of these data the authors recommended a shortened instrument consisting of the PPI, the Analgesic Score, 7 of the 17 LASAS (pain, physical activity, fatigue, appetite, family-marriage relationships, mood and overall well-being) plus an additional LASAS for constipation. An additional LASAS labelled 'passing urine' was added subsequently.

In the second study the the PPI, pain LASAS and Analgesic Score were used to assess the effects of intravenous mitoxantrone chemotherapy plus low-dose oral prednisone assessed in 27 men with symptomatic advanced HRPC. Response was defined as a 2 point or greater improvement in the PPI with no increase in analgesic score or a ≥ 50% reduction in analgesic score with no increase in PPI, both maintained for at least 2 consecutive assessments at least 3 weeks apart.

1.5.2 Description of the PROSOQLI (see appendix 1)

Each LASAS consists of a label (pain, physical activity, fatigue, appetite, constipation, family/marriage relationships, mood, passing urine, and overall well-being (How do you feel?) and an ungraduated, horizontal 10 cm line with verbal anchors at either end. The anchor signifying the worst symptoms, function or overall well-being is on the left. The lines are measured from left to right so that a score of 10 represents best function or quality-of-life and least symptoms. The LASAS for urinary symptoms was added to the PROSOQLI midway through the randomised trial. For the PPI scale, subjects circle
a number from 0-5 labelled with a pain descriptor (0 = no pain, 1 = mild pain, 2 = discomforting pain, 3 = distressing pain, 4 = horrible pain, 5 = excruciating pain). The time frame for both the LASAS and the PPI is the past 24 hours.

Analgesic intake was recorded daily by the patient on a diary card. The Analgesic Score is calculated by multiplying the average daily intake of each drug during the last 7 days by an arbitrary score per 'standard' tablet or dose. A score of 1 is given for every 'standard' dose of non-narcotic analgesic (acetaminophen 325mg, aspirin 325mg, ibuprofen 200mg, indomethacin 25mg, naproxen 250mg) or 'weak' mixed analgesic (acetaminophen 325mg with codeine 8mg). A score of 2 is given for each 'standard' dose of narcotic analgesic (morphine 10mg, hydromorphone 2mg, codeine 60mg) or 'stronger' mixed analgesic (codeine 30mg with acetaminophen 325mg, oxycodone 5mg with either acetaminophen 325mg or aspirin 325mg). Thus a patient taking 100 mg of oral morphine per day or 10 tylenol #3 per day would have an analgesic score of 20. These scores do not reflect analgesic potency accurately, particularly across classes: the variability of individual responses to different classes, drugs, and doses precludes meaningful comparisons across individuals. The aim of the analgesic score is to assess changes in analgesic use over time in an individual rather than to compare analgesic use between individuals.
1.6 The EORTC instrument (EORTCI)

1.6.1 Development and prior validation of the EORTC QLQ-C30

The EORTC QLQ-C30 was developed to evaluate the quality-of-life of cancer patients participating in international clinical trials. This 30-item questionnaire is designed to address a range of quality-of-life issues relevant to cancer patients in general and is intended to be the core of a modular approach in which supplementary questionnaires ("modules") are used to address additional aspects of quality-of-life which are of interest in a particular study. The theoretical and methodologic framework, influenced strongly by the work of Ware et al., is of HRQL as a multidimensional abstract concept reflecting physical, psychological and social aspects.

The first generation questionnaire (QLQ-C36) was field-tested in 537 patients with lung cancer. The second generation questionnaire (QLQ-C30 as used in this study) was field-tested in a second sample of 305 patients with lung cancer. Assessments of internal consistency and correlations between scales supported structure of the questionnaire, whilst comparisons against independent measures of performance status, weight loss, and treatment toxicity support its discriminative validity. These studies provide strong support for the psychometric validity of the core questionnaire in patients with non-resectable lung cancer. The restriction to this single clinical condition limits the generalisability of the results to other cancers. However, the populations where actually quite heterogeneous in terms of disease extent, symptoms and performance status. This heterogeneity produces higher correlations and better discrimination than would be the case with a more homogenous population typical of a controlled clinical trial.

Subsequent studies in large groups of patients with miscellaneous types of cancers
have lent further support to the validity of the EORTC-QLQ-C30.

Ringdal et al assessed the scaling structure of the EORTC QLQ-C30 in 177 people with a variety of types of cancer and found support for all scales except cognitive function and role function, neither of which are considered in the current thesis. Factor analysis revealed two factors, one correlating strongly with physical function and symptoms, the other correlating most strongly with emotional, social and cognitive functioning. The report does not describe the patient population clearly and fails to acknowledge the influence of the study population on the observed correlations.

Niezogoda et al assessed the convergent and divergent validity of the EORTC QLQ-C30 with the Sickness Impact Profile, the McGill Pain Questionnaire, the General Health Questionnaire and the CARES in 96 people with miscellaneous types of cancers. The results provided strong support for the scales assessing physical function, pain, and fatigue, and moderate support for the scale assessing emotional function. The scale for social function was strongly correlated with scales assessing physical function (0.41 to 0.74), reflecting the wording of the (social function) items which specify limitations due to physical condition or medical treatment.

Osoba et al assessed construct validity in 535 people participating in a randomised trial of antiemetics for nausea and vomiting induced by chemotherapy. The population was very heterogeneous in terms of the types of cancer and the extent of disease including some patients being treated with curative intent and others with palliative intent. The intended grouping of items into scales was supported by factor analysis, item-domain correlations and domain-domain correlations. Discriminative validity was supported by comparisons against performance status and the presence or absence of metastatic disease.
This study extends support for the QLQ-C30's validity to a broader range of cancers, although not prostate cancer. The heterogeneity of patients in this study will exaggerate indices of convergent and discriminative validity if the intended target population is more homogeneous.

Hjermstad et al assessed test-retest reliability in 190 out-patients with a variety of cancers by administering the EORTC QLQ-C30 on two occasions, four days apart. They used Pearson's and Spearman's correlation coefficients, indices of association, to express reliability. The correlations ranged from 0.63 to 0.91 indicating strong associations between the two ratings. The mean change scores between the two assessments were claimed to be low suggesting that the reported correlation coefficients should provide a reasonable approximation to the preferred index of reliability, the intraclass correlation coefficient. The heterogeneity of the population in terms of their diagnoses would tend to inflate the estimates of reliability. The majority of subjects were in a stable phase of their disease: most (86%) were in clinical remission and none had received treatment within the last 3 months. This would also tend to give estimates of reliability higher than those expected in patients receiving treatment for active and symptomatic disease. Nonetheless, this study provides moderate support for the test-retest reliability of the instrument in a stable population.

Published evidence for the responsiveness of the QLQ-C30 is restricted to comparisons with changes in performance status and with changes due to the acute toxicity of chemotherapy. Aaronson et al reported responsiveness of the QLQ-C30 to changes in ECOG performance status. The ECOG performance status scale has only 5 categories (see Appendix 5) and 93% of patients fell into the best 3 categories, therefore changes in
ECOG performance status represent large changes in clinical status. Thus, the ability to detect these large changes in performance status represents only weak evidence of responsiveness. Osoba et al reported the responsiveness of the scales for physical function, role function, social function, global quality of life, nausea, vomiting and fatigue to changes in quality of life associated with the administration of moderately and severely emetogenic chemotherapy. This study provides strong evidence of the instrument’s responsiveness to the changes associated with the toxicity of emetogenic chemotherapy. It provides no evidence of the responsiveness of the instrument to either positive changes due to treatment or changes in either direction due to disease. Additional studies assessing responsiveness have been presented in abstract form. A 32-item version of the EORTC core quality of life questionnaire was used in the phase 2 study of mitoxantrone plus prednisone described in sections 1.2.4 and 1.5.1. This study demonstrated changes in pain, anorexia, social function, and emotional function associated with treatment, but is difficult to interpret in the absence of a control group or an alternative measure.

1.6.2 The QLM-P14

The QLM-P14 is an expanded version of a specific module developed by David Osoba for the phase 2 study of mitoxantrone plus prednisone according to the guidelines of the EORTC Committee for Quality of Life Research. Only the urinary items of the QLM-P14, added midway through the randomised trial, are considered in the current study.

1.6.3 Description of the EORTCI (see appendices 2 and 3)

The EORTC QLQ-C30 and QLM-P14 are based predominantly on items with a 4-point ordinal response format where patients choose from: 'not at all', 'a little', 'quite a bit' or 'very much' in response to a series of questions. The EORTC QLQ-C30 includes multi-
item Likert scales for emotional function, social function, pain and fatigue; single-item scales for anorexia and constipation; a Likert scale of 5 items with a yes/no response format for physical function; and 2 items with 7-point numeric responses for global quality-of-life. The QLM-P14 includes 2 items addressing urinary symptoms which were added midway through the randomized trial. Both the EORTC QLQ-C30 and the QLM-P14 include additional domains not covered by the PROSQOLI and not discussed further in this thesis. In this thesis the term EORTC refers only to those scales which have been matched to the scales of the PROSQOLI, namely the QLQ-C30 scales for physical function, emotional function, social function, pain, fatigue, appetite, and constipation and the QLM-P14 scale for urinary symptoms. Appendix 4 illustrates the components of the EORTC and compares it with the PROSQOLI.
1.7 The conventional measures

1.7.1 Performance status

Performance status is an index of physical functional capacity reflecting the ability to work, move about, and care for oneself (see Appendix 5). In the present study performance status was assessed with the 5-point ordinal index developed by Zubrod et al for the Eastern Cooperative Oncology Group (ECOG) which has gained widespread acceptance in cancer research. A Medline search of 5 years of published cancer research in which the assessment of performance status was mentioned in the abstract revealed that 50% of studies used the ECOG index and the remaining 50% used the Karnofsky index. In the present study, performance status was measured by the study nurse or physician at baseline as a variable for stratified randomisation. Two studies have addressed the cross-sectional interobserver reliability of the ECOG performance status scale. In an Italian study, 209 consecutive cancer patients entering a university department of medical oncology were rated independently by two medical oncologists. Interobserver reliability assessed by the quadratically-weighted kappa statistic was excellent (0.91). In a Danish study, 3 observers independently rated 100 consecutive in-patients. Reliability assessed with a kappa statistic was lower at 0.44, although it is not clear which form of the kappa statistic was used. Reliability was excellent in both studies when performance status was scored as a dichotomous variable (0-2 versus 3-4). Numerous studies have identified performance status as a strong predictor of survival duration in patients with cancer.
1.7.2 Prostatic acid phosphatase (PAP) & prostate specific antigen (PSA)

PAP and PSA are produced predominantly by prostate cells, both normal and cancerous. PAP is more specific for prostate cancer than PSA: elevated blood levels of PAP are rare in men without prostate cancer whereas elevated levels of PSA can occur in men with benign abnormalities of the prostate. PSA is more sensitive for prostate cancer than PAP: more men with prostate cancer have elevated blood levels of PSA than of PAP. Since both PAP and PSA are produced by prostate cancer cells, their absolute blood levels are thought to reflect tumour bulk; although this is probably true across the spectrum of prostate cancer it is not clear that this it is true for men with advanced prostate cancer. Blood levels of both PAP and PSA have been assessed as outcomes in advanced prostate cancer; several groups have recommended that the percentage drop in PSA be used as an outcome measure for clinical trials in advanced prostate cancer. Blood levels of PAP are expressed as the ratio of the measured level to the upper limit of normal for the particular laboratory; levels ≤ 1 are normal. For PSA the same assay was used in all centres so raw blood levels are reported; although blood levels vary according to age, levels less than 4 mg/l are reported as normal.

1.7.3 Alkaline Phosphatase (ALP), hemoglobin & creatinine

In advanced prostate cancer, elevated blood levels of ALP are usually associated with bone metastases but the relationship is complicated because ALP levels are also elevated in association with bone repair. Elevated levels of ALP may also occur in association with liver damage but liver metastases are much less common than bone metastases in prostate cancer (visceral metastases in 4% compared with bone metastases in 96% of patients in the randomised trial). To adjust for differences between laboratories,
blood levels of ALP are expressed as the ratio of the measured level to the upper limit of normal for the particular laboratory: levels ≤ 1 are normal.

Hemoglobin is produced by red blood cells. In men with advanced prostate cancer depressed blood levels of hemoglobin usually reflect bone marrow damage by cancer, treatment or both. Abnormally low blood levels of hemoglobin (anemia) are often associated with fatigue and weakness.

Blood levels of creatinine are indicative of kidney function. In men with advanced prostate cancer elevated levels of creatinine are usually associated with kidney damage caused by urinary obstruction. A near normal creatinine level was an eligibility criterion for the randomised trial so the range of creatinine concentrations in this trial is narrow.

1.8 Administration of indices

The PROSQOLI and the EORTC instrument were administered at each 3 weekly clinic visit. Patients completed the questionnaires without assistance after an initial explanation by a study nurse. The questionnaires were presented in the same order to all patients: the PROSQOLI first followed by the EORTC instrument. Patients kept a daily diary of their analgesic intake. The Analgesic Score was based on the last seven days prior to the clinic visit.

Performance status was rated by a physician or study nurse at baseline (see appendix 5). Pretreatment blood tests included alkaline phosphatase (ALP), prostatic acid phosphatase (PAP), prostate specific antigen (PSA), creatinine, and complete blood count (including hemoglobin); these were to be repeated every 3 weeks. Pretreatment imaging included a radioisotope bone scan, and x-rays of the chest, pelvis and painful bony sites; these were to be repeated every 12 weeks. Baseline serum PSA levels were available for
only 134 of the 161 patients because this test was not performed at all centres when the study began in 1989. Survival time was measured from the date of randomization.

1.9 Summary

The PROSQOLI is designed to be a pragmatic outcome measure for clinical trials of systemic treatments for symptomatic men with advanced hormone-resistant prostate cancer. It was developed and refined in the course of separate single-arm studies which assessed the activity of prednisone alone and prednisone plus mitoxantrone in this patient population. The PROSQOLI was the primary measure of HRQL in a recently completed randomised trial comparing these two treatments. HRQL was also assessed with the EORTCI which was judged to be the best available alternative instrument when the randomised trial was designed in 1989. Conventional clinical outcome variables such as blood tests and performance status were also assessed in the trial. This trial demonstrated a clinically important and statistically significant benefit associated with treatment with mitoxantrone plus prednisone. The objective of the research reported in this thesis is to assess the cross-sectional validity of the PROSQOLI using the baseline data from the randomised trial.
2 VALIDATION IN GENERAL

2.1 Rationale

In order to achieve its intended status as a scientific instrument, the PROSQOLI must be considered valid by its intended audience. Validity is much like quality of life: it is a multidimensional, abstract concept; most people feel familiar with it and most people agree that it is important; however on closer examination, it seems to mean quite different things to different people. The dictionary definition of validity includes soundness, justness, well-foundedness and cogency. It is not surprising then, as Nunnally points out, that:

"... the term has considerable "surplus meaning," in that it implies all things good about a measuring instrument rather than explicitly indicating the standards by which measuring instruments must be judged."

Nunnally and others, use the term validity to refer to the degree to which an instrument measures what it is supposed to measure. This definition emphasises that validity is not an all or none phenomenon, that it depends on hypotheses rather than facts, and that it may be assessed in more than one way. Ultimately, validity is in the eye of the beholder: the aim of this thesis is to provide sufficient data for the reader to make an informed judgement.

2.2 Empirical validity

Empirical validity refers to those aspects of validity which are assessed with data arising from use of the instrument. Face validity (the extent to which an index looks as if it is measuring what it is supposed to measure), content validity (the extent to which an index covers those aspects thought to be important) and other aspects of what Feinstein calls sensibility are not considered in this thesis.
2.3 Cross-sectional validity

Cross-sectional refers to data gathered with the PROSQOLI at a single time point. Kirshner and Guyatt suggested that health indices be assessed in terms of their stated purpose and proposed a trichotomy of purposes: discrimination, prediction, and evaluation. Discrimination refers to distinguishing between individuals with true differences in the attribute being measured, while prediction refers to distinguishing between individuals with differences in subsequent outcomes. Discrimination and prediction are both cross-sectional in that they can be assessed with a single administration of an index. The traditional psychometric concept of validity is cross-sectional reflecting the predominantly cross-sectional purposes of discrimination and prediction for which psychological indices have been developed. The aspects of validity considered in this thesis are all cross-sectional.

Evaluation refers to distinguishing differences within an individual over time. Evaluation is longitudinal in that it can only be assessed with repeated administrations of an index. Responsiveness, the ability to detect change over time, is the crucial measurement property of an evaluative instrument. Since the PROSQOLI is an evaluative instrument, responsiveness is a component of validity as we have defined it: the PROSQOLI was developed to detect changes associated with treatment and disease: therefore, its ability to detect change over time is crucial to its validity. The measurement properties which are optimal for an index intended to detect differences within an individual over time may differ from those which are optimal for an index intended to detect differences between individuals at a single time point. Guyatt et al point out that under certain circumstances these requirements may be conflicting and suggest that the
concepts of cross-sectional validity and responsiveness should therefore be kept
distinct.\textsuperscript{63,65,66} Hays et al conceptualise responsiveness as an aspect of validity rather than
as being distinct from it.\textsuperscript{67}

Whilst it is clear that the optimal characteristics of an index may differ depending
on whether its primary purpose is discrimination, prediction, or evaluation and that under
some circumstances these requirements may be conflicting, it also seems clear that both
cross-sectional validity and responsiveness are desirable properties. Cross-sectional validity
is a more basic concept and has therefore been addressed first. Responsiveness is crucial
to the validity of the PROSOQLI and is considered in a separate report.

2.4 Construct validity

In the absence of a gold-standard against which the validity of an instrument can
be judged directly a variety of indirect methods are used. In general, these indirect
methods depend on hypotheses about how an instrument should behave given the theory
on which the instrument is based. Specifically, hypotheses are made about the construct
of interest, and these hypotheses are tested empirically with the instrument which is
supposed to measure that construct. Empirical support for the hypotheses lends indirect
or inferential support for the validity of the instrument: that the instrument is measuring
what it is supposed to be measuring. For example, to test the discriminative validity of the
global HRQL scales, their ability to detect differences between groups classified according
to performance status was assessed. This assessment presupposes that people with
different levels of performance status have differences in self-assessed global quality of life,
that the scale used to measure performance status is valid and was rated reliably. The
assessment of validity is as much a test of the assumptions as it is a test of the scale under
2.5 Clinimetric and psychometric approaches to validation

Health-related quality of life research represents an intersection of two scientific traditions which can be labelled as psychosocial and biomedical. These traditions are characterised by differing goals, techniques, and criteria for success reflecting the backgrounds and interests of their proponents. The attributes of HRQL are typical of psychosocial research in that they are subjective, multidimensional and intangible. Psychometrics is the science of measuring mental capacities or processes; it includes rigorous and sophisticated methods for the development and validation of instruments suited to the needs of psychosocial researchers. The psychometric approach has provided a framework for the development of indices to assess subjective phenomena in biomedical research. The goals of biomedical research, however, are distinct from the goals of psychosocial research.

Feinstein coined the term clinimetrics for the discipline concerned with measuring clinical phenomena with the aim of formalising rigorous methods for the development and validation of instruments suited to the needs of biomedical researchers. Wright and Feinstein have highlighted the similarities and differences between the psychometric and clinimetric approaches emphasising that the strengths and weaknesses of the approaches are relative and dependent on the intended purpose for an index. These distinctions are relevant to the assessment of the validity of the PROSQOLI.

The PROSQOLI is a clinimetric index. It was developed by and for biomedical researchers using clinical experience to select heterogeneous domains assessed with single items. The assessment of homogeneity for multi-item scales, a key aspect of psychometric
validity, is not applicable to the PROSQOLI. The domains of the PROSQOLI are heterogeneous and distinct but relationships between certain domains are expected. The assessment of convergent validity, another fundamental psychometric concept, is highly applicable to the PROSQOLI.

The ability of the PROSQOLI to predict survival duration is used as a test of its validity even though it was not designed to predict survival duration. The psychometric and clinimetric concepts of predictive validity are quite different. The classical psychometric concept of predictive validity refers to the ability of an instrument to predict the outcome it is designed to measure. For example, the predictive ability of an instrument designed to measure academic potential is assessed by its ability to predict subsequent academic performance; accurate predictions support the contention that the index is measuring academic potential. The clinimetric concept of predictive validity refers to the ability of an instrument to predict a subsequent outcome of interest. For example, the predictive validity of an index designed to predict survival duration is assessed by its ability to predict survival duration; however, accurate predictions do not support the contention that the instrument is measuring survival duration.

In this thesis both psychometric and clinimetric terminology and techniques are used, reflecting the intersection of the psychosocial and biomedical domains in HRQL research. Further discussion of these distinctions is left to the sections on distributional, convergent, discriminative and predictive validity.
2.6 Exploratory versus confirmatory analysis

The terms exploratory analysis and confirmatory analysis, as popularised by John W. Tukey, refer to differing aspects of data analysis. "Exploratory analysis is about looking at data to see what it seems to say...", whilst confirmatory analysis is about assessing the precision of inferences drawn from the data. Exploratory analysis includes description, summarisation, and hypothesis generation, whist confirmatory analysis includes tests of clearly specified a priori hypotheses. The analyses in this thesis are neither purely exploratory nor purely confirmatory. There are many equally important hypotheses under scrutiny most of which are best viewed comparatively, and none of which provide a clear criterion for acceptance or rejection. Rigour is afforded by a priori specification of the hypotheses, methods, and criteria and by acknowledging that there are elements of exploration and confirmation in each analysis.

2.7 P-values

P-values are used throughout this thesis to indicate levels of support for individual sub-hypotheses, the lower the p-value the greater the support; they are not intended as sharp criteria for proof or rejection. All p-values are two sided. No adjustments have been made for multiple comparisons. All hypotheses tested are evident in this report and were prespecified. P-values have been interpreted as continuous, that is without sharp demarcations at arbitrary levels, and conservatively; thus a p-value around .05 is regarded as providing tentative rather than definitive support for a hypothesis. In general, p-values have been interpreted in a comparative rather than an absolute sense. That is, they have been used to compare the levels of support across domains and instruments rather than to indicate acceptance or rejection of any particular hypothesis. The final decision
regarding what constitutes "reasonable doubt" is left to the reader.

2.8 Validation by application

This study, and those in which the PROSQOLI was developed, are examples of validation by application.63 Assessment of the validity of the PROSQOLI was specified as an objective in the protocol for the randomised trial and was the main reason for including two measures of HRQL, the other being corroboration of the results. The specific details of this validation study - the goals, methods, and criteria for interpretation - were written by Martin Stockler (MS), in collaboration with Ian Tannock (IT) and Pamela Goodwin (PG), during the last year of the trial and prior to examination of the data.
3 DESCRIPTIVE DATA

3.1 Introduction and aims

In this chapter the dataset is examined in order to learn about both the patients and the indices used to describe them.

3.2 Methods

3.2.1 Preparation of the data

The data used in this analysis are the baseline data from a randomised trial. The baseline data account for 161 of a total of 1267 rows of data - 1 row per patient per visit. Due to an error in photocopying, about 30% (361/1267) of the linear analog scales were only 9 cm long. This error was detected prior to the database being closed. A correction factor of 1.11 has been applied to all the short LASAS scores. The distributions and primary results were not different according to whether the scales were long or short.

The data from the completed forms were double-entered into a SAS database managed by the pharmaceutical company who manufacture mitoxantrone. After extensive quality control including range checks and examination of random records the database was closed and exported in an IBM-PC compatible format by the pharmaceutical company. All subsequent data-management and all analyses reported in this thesis were performed by Martin Stockler using the Statistical Package for Interactive Data Analysis (SPIDA).70

The baseline data were examined in detail to ascertain their integrity. One patient was found to have 2 sets of baseline data - the original forms were checked and the database corrected. Several inconsistencies were detected in the dates, all due to transposition of the day and month.
3.2.2 General descriptive methods

The distributions of the patient-based data are illustrated with frequency histograms and summary statistics. Non-parametric summary statistics are emphasised throughout the thesis since the distributions were rarely normal (see below), however parametric statistics are included where they add information, for example to facilitate sample size estimates.

3.2.3 Compliance with the patient-based measures

Compliance is described in terms of the proportions of questionnaires and individual domains completed with all randomised patients in the denominator. For the multi-item Likert scales of the EORTC instrument, any scale with a missing item was regarded as missing. Similarly, any LASAS without a mark was scored as missing.

3.2.4 Transformation of the patient-based measures

All HRQL scores were transformed linearly to a range from 0 to 100, where 0 represents worst quality of life (poorest function and most severe symptoms) and 100 represents best quality of life (highest function and least severe symptoms). The main advantage is that comparisons between the two instruments are simplified by expressing them in the same metric: signs and sizes of correlations, regression coefficients and differences between groups will have the same meaning across instruments and scales. The disadvantage is that for the symptom scales higher scores represent less symptoms. This choice is arbitrary and consistent with the LASAS of the PROSQOLI (the left end is always the worst) and the functional and global scales of the EORTC instrument, but opposite to the symptom scales of the EORTC instrument.

For the histograms the LASAS scores are rounded to the nearest 10 (i.e. measured in centimetres) whereas the histograms for the EORTCI scores and the PPI show the actual
actual response categories.

3.2.5 Non-parametric statistics

Non-parametric statistics have been used for all but the multivariable analyses. This choice was specified \textit{a priori} based on the anticipated distributions, controversy regarding the interval nature of the measures, the desire to limit assumptions, and for the sake of consistency.

Distributions of scores from HRQL indices are rarely symmetrical and almost never Gaussian; the data from this trial are typical. Although the attributes being measured are continuous, there is considerable controversy regarding the \textit{level of measurement} provided by the types of response scales used, namely, linear analog and Likert scales. The term level of measurement refers to the scale on which a measurement is expressed. The response scale may be categorical, where there is no logical ranking of levels (e.g. eye colour); ordinal where the levels can be ranked but the distance between levels is arbitrary (e.g. performance status); or dimensional where the intervals between adjacent categories are uniform (e.g. hemoglobin concentration). Single-item 4-point Likert scales such as those used to assess appetite and constipation in the EORTCI are clearly ordinal in nature. Feinstein regards both linear analog and multi-item Likert scales as quasi-dimensional\textsuperscript{63} while Streiner and Norman argue that,

"from a pragmatic viewpoint, it appears that under most circumstances, unless the distribution of scores is severely skewed, one can analyse data from rating scales as if they were interval without introducing severe bias."\textsuperscript{62}

Since the response scales are at least ordinal, non-parametric rank methods have been used in most instances. This choice is consistent with the analysis of the primary results of the
randomised trial and therefore reflects the general principle of assessing validity under circumstances as close as possible to those of intended use. Parametric summary statistics, such as the mean and standard deviation, have been included for the patient-based measures for the sake of completeness and to facilitate sample size calculations.
3.3 Findings

3.3.1 Description of patients

Baseline demographic and clinical details of the 161 men are summarised in Tables 1 and 2. All the men were symptomatic with advanced prostate cancer which had progressed after at least one hormonal therapy. Half the men were 63 to 74 years old, had an ECOG performance status of 1 or 2, rated their pain as mild or discomforting, and were receiving the "equivalent" of 8-25mg of oral morphine every 4 hours. Nearly all had demonstrable bony metastases, and about one in 5 were known to have lymph node metastases. Most had highly elevated serum concentrations of prostate specific antigen (PSA) (median = 176 ng/ml, upper limit of normal = 4ng/ml) and moderately elevated levels of prostatic acid phosphatase (PAP). About one in five of the men were anemic (hemoglobin < 100 g/l).

3.3.2 Conventional measures

3.3.2.1 Compliance with conventional measures

Compliance with the collection of the conventional measures was generally good with over 98% of the patients having data on performance status, time since diagnosis, hemoglobin level and the presence or absence of bony metastases. Baseline levels of serum PSA were available on only 83% of subjects since this test was not available in all participating centres when the study began.

3.3.2.2 Distributions of conventional measures

The distributions of the blood tests ALP, PAP, PSA and creatinine were positively skewed and all were made more symmetric by a logarithmic transformation. The distribution of the analgesic scores was also very skewed to the right.
Table 1. Baseline descriptive data: continuous conventional measures.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Median</th>
<th>Interquartile Range</th>
<th>Range</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>68</td>
<td>63 - 74</td>
<td>44 - 86</td>
<td>160</td>
</tr>
<tr>
<td>Years since diagnosis</td>
<td>3</td>
<td>1.5 - 4.9</td>
<td>0.2 - 16</td>
<td>161</td>
</tr>
<tr>
<td>Prostate Specific Antigen (ng/ml)</td>
<td>176</td>
<td>53 - 564</td>
<td>0.2 - 6290</td>
<td>134</td>
</tr>
<tr>
<td>Prostatic Acid Phosphatase*</td>
<td>4.3</td>
<td>1.1 - 17</td>
<td>0.1 - 620</td>
<td>139</td>
</tr>
<tr>
<td>Alkaline Phosphatase*</td>
<td>2.3</td>
<td>1.3 - 5.2</td>
<td>0.1 - 37</td>
<td>153</td>
</tr>
<tr>
<td>Hemoglobin (g/l)</td>
<td>119</td>
<td>105 - 128</td>
<td>74 - 150</td>
<td>158</td>
</tr>
</tbody>
</table>

* ratio of reported level to upper limit of normal (uln) for that laboratory (i.e. 1 = uln)

Table 2. Baseline descriptive data: categorical conventional measures.

<table>
<thead>
<tr>
<th>Performance Status (ECOG)</th>
<th>Percentage</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>5%</td>
<td>8</td>
</tr>
<tr>
<td>1</td>
<td>58%</td>
<td>92</td>
</tr>
<tr>
<td>2</td>
<td>27%</td>
<td>43</td>
</tr>
<tr>
<td>3</td>
<td>10%</td>
<td>16</td>
</tr>
<tr>
<td>missing</td>
<td>1%</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metastases present</th>
<th>Percentage</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bone Metastases present</td>
<td>96%</td>
<td>155</td>
</tr>
<tr>
<td>Lymph Node Metastases</td>
<td>21%</td>
<td>33</td>
</tr>
<tr>
<td>Visceral Metastases</td>
<td>4%</td>
<td>6</td>
</tr>
</tbody>
</table>

Total | 161
3.3.3 Patient-based measures

Descriptive statistics (Table 3) and frequency histograms (Figure 1) for the distributions of baseline scores for the two patient-based instruments are shown together for reference. These summary statistics are described and discussed in detail in the following sections.

3.3.3.1 Compliance with patient-based measures

Baseline questionnaires were completed by all but one of the 161 randomized patients. Most patients answered every question - the lowest response rate for the PROSQOLI was 98% (157/161) for the family/marriage relationships LASAS. Ability to complete the HRQL questionnaires was an eligibility criterion for the trial.

The PROSQOLI LASAS were completed at least as often as the EORTC Likert scales suggesting that this group of elderly men on narcotic analgesics did not find the linear analog response format more difficult to complete. However, box plots of the LASAS divided according to the scores for the analogous EORTC scale identified extreme outliers for most scales. For example, two subjects answered "very much" for the two Likert scales assessing pain ("Have you had pain?", and "Did pain interfere with your daily activities?"), but rated their pain as being very mild with the LASAS (7 and 11 mm from the end labelled "no pain at all"). It is likely that these aberrant ratings reflect mistakes in answering the LASAS rather than the Likert scales. All scales were scored as completed; no attempt was made to "correct" responses which appeared inconsistent.
Table 3. Baseline descriptive data for the patient-based measures. All scores have been linearly transformed such that 0 is the worst possible score and 100 is the best possible score.

<table>
<thead>
<tr>
<th>Scales</th>
<th>Size</th>
<th>Mean</th>
<th>SD</th>
<th>Min LQ</th>
<th>Median</th>
<th>UQ</th>
<th>Max</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Omnibus test for normality</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROSQLI Scales</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 Pain LASAS</td>
<td>158</td>
<td>62</td>
<td>23</td>
<td>3</td>
<td>44</td>
<td>63</td>
<td>82</td>
<td>100</td>
<td>-0.3</td>
<td>2.1</td>
</tr>
<tr>
<td>2 Fatigue</td>
<td>159</td>
<td>49</td>
<td>25</td>
<td>0</td>
<td>32</td>
<td>46</td>
<td>69</td>
<td>100</td>
<td>0.2</td>
<td>2.3</td>
</tr>
<tr>
<td>3 Appetite</td>
<td>160</td>
<td>66</td>
<td>31</td>
<td>0</td>
<td>41</td>
<td>75</td>
<td>96</td>
<td>100</td>
<td>-0.5</td>
<td>1.9</td>
</tr>
<tr>
<td>4 Constipation</td>
<td>160</td>
<td>58</td>
<td>31</td>
<td>0</td>
<td>30</td>
<td>58</td>
<td>87</td>
<td>100</td>
<td>-0.2</td>
<td>1.8</td>
</tr>
<tr>
<td>5 Passing Urine</td>
<td>85</td>
<td>78</td>
<td>26</td>
<td>0</td>
<td>70</td>
<td>89</td>
<td>97</td>
<td>100</td>
<td>-1.4</td>
<td>3.9</td>
</tr>
<tr>
<td>6 Physical Activity</td>
<td>160</td>
<td>66</td>
<td>23</td>
<td>7</td>
<td>49</td>
<td>69</td>
<td>84</td>
<td>100</td>
<td>-0.4</td>
<td>2.4</td>
</tr>
<tr>
<td>7 Mood</td>
<td>160</td>
<td>71</td>
<td>23</td>
<td>0</td>
<td>53</td>
<td>76</td>
<td>91</td>
<td>100</td>
<td>-0.7</td>
<td>2.7</td>
</tr>
<tr>
<td>8 Family/Marriage R'ships</td>
<td>157</td>
<td>91</td>
<td>16</td>
<td>1</td>
<td>90</td>
<td>96</td>
<td>100</td>
<td>100</td>
<td>-3.3</td>
<td>15.4</td>
</tr>
<tr>
<td>9 Overall Well-being</td>
<td>160</td>
<td>62</td>
<td>23</td>
<td>0</td>
<td>47</td>
<td>61</td>
<td>81</td>
<td>100</td>
<td>-0.5</td>
<td>2.9</td>
</tr>
<tr>
<td>10 Pain PPI</td>
<td>161</td>
<td>61</td>
<td>18</td>
<td>20</td>
<td>60</td>
<td>60</td>
<td>80</td>
<td>100</td>
<td>-0.5</td>
<td>2.7</td>
</tr>
<tr>
<td>EORTC Scales</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11 Pain Likert</td>
<td>158</td>
<td>49</td>
<td>26</td>
<td>0</td>
<td>33</td>
<td>50</td>
<td>67</td>
<td>100</td>
<td>-0.0</td>
<td>2.1</td>
</tr>
<tr>
<td>12 Fatigue</td>
<td>153</td>
<td>49</td>
<td>22</td>
<td>0</td>
<td>33</td>
<td>44</td>
<td>67</td>
<td>100</td>
<td>0.0</td>
<td>2.8</td>
</tr>
<tr>
<td>13 Appetite</td>
<td>157</td>
<td>64</td>
<td>34</td>
<td>0</td>
<td>33</td>
<td>67</td>
<td>100</td>
<td>100</td>
<td>-0.6</td>
<td>2.2</td>
</tr>
<tr>
<td>14 Constipation</td>
<td>160</td>
<td>57</td>
<td>33</td>
<td>0</td>
<td>33</td>
<td>67</td>
<td>67</td>
<td>100</td>
<td>-0.4</td>
<td>2.1</td>
</tr>
<tr>
<td>15 Urinary Symptoms</td>
<td>82</td>
<td>72</td>
<td>21</td>
<td>0</td>
<td>67</td>
<td>67</td>
<td>83</td>
<td>100</td>
<td>-1.0</td>
<td>4.1</td>
</tr>
<tr>
<td>16 Physical Function</td>
<td>155</td>
<td>52</td>
<td>26</td>
<td>0</td>
<td>40</td>
<td>60</td>
<td>60</td>
<td>100</td>
<td>-0.0</td>
<td>2.6</td>
</tr>
<tr>
<td>17 Emotional Function</td>
<td>155</td>
<td>69</td>
<td>20</td>
<td>17</td>
<td>50</td>
<td>67</td>
<td>83</td>
<td>100</td>
<td>-0.2</td>
<td>2.3</td>
</tr>
<tr>
<td>18 Social Function</td>
<td>157</td>
<td>59</td>
<td>28</td>
<td>0</td>
<td>33</td>
<td>67</td>
<td>83</td>
<td>100</td>
<td>-0.3</td>
<td>2.3</td>
</tr>
<tr>
<td>19 Global Quality of Life</td>
<td>159</td>
<td>45</td>
<td>21</td>
<td>0</td>
<td>33</td>
<td>50</td>
<td>58</td>
<td>100</td>
<td>0.1</td>
<td>3.0</td>
</tr>
</tbody>
</table>

Size = number of completed scales. SD = sample standard deviation. Min = minimum value. LQ = lower quartile. UQ = upper quartile. Max = maximum value. Skewness = 0 for a normal distribution. Kurtosis = 3 for a normal distribution. $X^2$ = chi-square on 2 degrees of freedom from the omnibus test for the null hypothesis that the values are normally distributed. p = the associated p-value - the lower the p-value, the stronger the evidence that the distribution is not normal.
Figure 1. Histograms of the baseline distributions for the patient-based measures (0 is worst and 100 is best for all scales).
3.3.3.2 Distributions of the patient-based measures

Figure 1 shows the histograms of the baseline scores for the PROSQOLI and the EORTCI. The scores for most of the scales spanned the whole response range. The main exceptions were the PROSQOLI PPI scale for which no subjects rated their pain as "excruciating" and the PROSQOLI LASAS for family/marriage relationships. The histograms also show that many of the scales have asymmetric distributions.

The distributions of the baseline scores from the PROSQOLI LASAS were all negatively skewed with the exception of the LASAS for fatigue (Table 3 and Figure 2). Skewness is a measure of the asymmetry of a distribution, and is conventionally defined as the standardised third moment: \( E(X-\mu)^3/\sigma^3 \), where \( E(.) \) is the expected value operator, \( X \) is a randomly selected member of the population, \( \mu \) is the population mean and \( \sigma \) is the population standard deviation. A skewness of zero indicates symmetry around the mean which is what is expected for a normal distribution. A negative value indicates skewness to the left (towards values smaller than the mean - worse scores for these quality of life measures) and a positive value indicates skewness to the right (towards values higher than the mean - better scores for these quality of life measures). In a distribution which is skewed to the left, the left (lower) tail is longer than the right (higher) tail, and the median is higher than (to the right of) the mean. The asymmetry was greatest for the LASAS for family/marriage relationships and passing urine (skewed highly to the left) and least for fatigue. The EORTC scales also tended to be skewed to the left although to a lesser extent than the analogous PROSQOLI scales. The scale for urinary symptoms was the most skewed of the EORTC scales. The EORTC scale for social function was much less skewed than the analogous PROSQOLI LASAS for family/marriage relationships.
Normality of the Baseline Distributions

Skewness

PROSQOLI Scales
Family/Marriage R'ships
Passing Urine
Mood
Appetite
Overall Well-being
Pain PPI
Physical Activity
Pain LASAS
Constipation
Fatigue

EORTC Scales
Urinary Symptoms
Appetite
Constipation
Social Function
Emotional Function
Pain Likert
Fatigue
Physical Function
Global Quality of Life

Skewness = 0 for a normal distribution

Figure 2. Skewness of the distributions of baseline scores for the patient-based measures.
Kurtosis is a measure of curvature or "peakedness", and is measured by the fourth standardised moment: $E(X-\mu)^4/\sigma^4$. The kurtosis is 3 for a normal distribution; values higher than 3 generally reflect a narrower, higher peak with more observations in the tails while values less than 3 generally reflect a broader, lower peak with fewer observations in the tails than expected for a normal distribution. The scales which were most skewed were also most kurtotic reflecting the large proportion of values in their lower tails (see Figure 3).
Normality of the Baseline Distributions

Kurtosis

PROSQOLI Scales
Constipation
Appetite
Pain LASAS
Fatigue
Physical Activity
Mood
Pain PPI
Overall Well-being
Passing Urine
Family/Marriage R’ships

0 3 6 9 12 15 18

Kurtosis = 3 for a normal distribution

EORTC Scales
Pain Likert
Constipation
Appetite
Emotional Function
Social Function
Physical Function
Fatigue
Global Quality of Life
Urinary Symptoms

Figure 3. Kurtosis of the distributions of baseline scores for the patient-based measures.
The omnibus test for normality is based on deviations of the skewness and kurtosis of a distribution from the values expected for a normal distribution (i.e. 0 and 3 respectively). This test revealed departures from normality for all of the PROSQOLI scales. These departures were greatest for the LASAS for family/marriage relationships and passing urine and least for the LASAS for overall well-being and the PPI scale for pain (see Figure 4). There were significant departures from normal for all of the EORTC scales except those for fatigue, global quality of life and physical function.
Normality of the Baseline Distributions
Omnibus Test of Normality

PROSQOLI Scales
- Overall Well-being
- Pain PPI
- Fatigue
- Physical Activity
- Mood
- Pain LASAS
- Passing Urine
- Appetite
- Constipation
- Family/Marriage R'ships

Chi-square with 2 degrees of freedom

EORTC Scales
- Fatigue
- Global Quality of Life
- Physical Function
- Emotional Function
- Social Function
- Urinary Symptoms
- Pain Likert
- Constipation
- Appetite

Chi-square with 2 degrees of freedom

Figure 4. Omnibus tests of normality for the distributions of baseline scores for the patient-based measures. Chi-square values of 6, 12, and 24 represent p-values of .05, .002, and $10^{-5}$ respectively.
The PROSQOLI was designed with the idea that pain is the dominant symptom in men with advanced hormone resistant prostate cancer. Figure 5 shows the PROSQOLI scales ranked according to their medians from worst to best. In fact the LASAS with the worst median score was that for fatigue, followed by constipation, overall well-being, pain and physical activity. The EORTC scales were ranked similarly: fatigue, pain, global quality of life and physical function (again from worst to best). It is not clear how much these rankings reflect the severity of the symptoms, their perceived importance, or subtleties in the wording of the items. A questionnaire asking patients to rate the importance of each symptom has been designed and will be administered in a subsequent study.
Median Scores for the Patient-Based Measures
Ranked from Worst to Best

PROSQOLI Scales
- Fatigue
- Constipation
- Pain PPI
- Overall Well-being
- Pain LASAS
- Physical Activity
- Appetite
- Mood
- Passing Urine
- Family/Marriage R'ships

EORTC Scales
- Fatigue
- Pain Likert
- Global Quality of Life
- Physical Function
- Appetite
- Constipation
- Urinary Symptoms
- Emotional Function
- Social Function

Figure 5. Ranking the domains according to the median scores for the patient-based scales.
The standard deviations are remarkably similar across both scales and instruments: (median: 23, interquartile range 21-28). Figure 6 shows this graphically with boxplots of two measures of spread, the standard deviation and the intertertile range, for the data scaled from 0 to 1 rather than 0 to 100. The intertertile range (67th percentile - 33rd percentile) is a non-parametric measure of spread which is approximately equal to the standard deviation for a normal distribution. The median value of the intertertile ranges is almost identical to that of the standard deviations but the spread of the intertertile ranges is greater reflecting the greater variability of this statistic. The consistency of spreads across domains, scales and instruments suggests a method effect: that the spread of the scores is largely determined by the constraints of the response scale. This suggests that an estimate of 30 (on a scale from 0 to 100) for the standard deviation would be a reasonable starting point for sample size calculations.

Figure 6. Boxplots of two measures of spread for the baseline distributions of the patient-based scores. For a normal distribution the intertertile range $\approx$ standard deviation.
3.4 Discussion

The distributions of baseline scores indicate possible problems with three of the scales from the PROSQOL: the ordinal adjectival scale of pain intensity (PPI) and the LASAS for family/marriage relationships and passing urine.

Ninety-three percent of the men had PPI scores of 1 (mild, 33%), 2 (discomforting, 42%), or 3 (distressing, 19%) while 6% rated their pain as 4 (horrible) and no patient rated their pain as 5 (excruciating). Furthermore, since pain was an eligibility criterion, the baseline scores were restricted to 4 categories. An additional intermediate category might improve the spread of the scores. This would be particularly important if the PPI performed poorly in tests of responsiveness.

The scores for the LASAS for family/marriage relationships were highly skewed with 86% of respondents scoring > 80. This ceiling effect in which a large proportion of respondents nominate the best score limits the ability of the scale both to discriminate between individuals and to detect improvements within individuals. This is likely to be due partly to the ambiguity of the label (Family/Marriage Relationships) and partly to the anchors: extremely bad → normal for me. Other workers have found judgemental questions for the assessment of social function to perform poorly.9,11 People appear to be reluctant to criticise their relationships and social supports. Ware and others have suggested that questions addressing the impact of illness and treatment on social function, or questions asking for quantitation of the amount of time spent with others, are preferable.9 The two items comprising the EORTC domain for social function follow this format and the spread of the scores is much better. A new LASAS scale has been devised to assess social function and is to be compared with the existing scale in a subsequent study.
The LASAS for passing urine is also highly skewed, with 66% of respondents scoring > 80. The use of the ambiguous term "normal" as the anchor for the better end of the scale is likely to be at least part of the explanation. The label "passing urine" may also be ambiguous since the range of possible urinary symptoms for men with advanced prostate cancer is wide and includes inability to start passing urine, dribbling, incontinence, urinary frequency, pain on passing urine and passing blood in the urine. The distribution of scores for the two-item EORTC domain for urinary symptoms is less skewed but shows much the same pattern. It is possible that men with advanced prostate cancer have less urinary symptoms than was anticipated, or else that they regard urinary symptoms as being less important than other symptoms. A new LASAS has been devised to assess urinary symptoms and is to be compared with the existing scale.

3.5 Conclusions

The baseline distributions support the validity of the PROSQOLI LASAS for pain, fatigue, appetite, constipation, physical activity, and mood. The distributions for family/marriage relationships and passing urine are highly skewed suggesting that these scales will provide limited information. The EORTC scales followed a similar pattern but were better spread in general, particularly the social function scale.
4 CONVERGENT VALIDITY

4.1 Introduction and Aims

Convergent validity refers to the degree to which an instrument behaves like another instrument thought to be measuring the same attributes. Bivariate relationships between the scales of the PROSQOLI and the EORTCI were assessed with a multitrait-multimethod correlation matrix. Multivariable linear regression was used to explore the relationships between the individual symptom and function scales and the global scales.

4.2 Methods

4.2.1 Bivariate relationships

4.2.1.1 The multitrait-multimethod matrix approach

The multitrait multimethod matrix (MTMMM) approach is a method for the assessment of convergent (and discriminant) validity. It involves the examination of the matrix of bivariate correlations between multiple methods (different scales) and multiple traits (different domains) and provides a framework within which to interpret the correlation coefficients.\(^72\) The main benefits of the MTMMM approach are that it provides a strategy for dealing with large and complicated correlation matrices, and that it encourages a priori specification of the expected or desired magnitudes of the correlations. The method was developed for the assessment of cross-sectional psychosocial measures aimed at discriminating between individuals within whom the differing traits are expected to be distinct and independent. In the present study, the population is much more homogeneous than is usual in psychosocial studies and moderate correlations are expected between related but distinct domains such as fatigue and physical activity.
4.2.1.2 Index of association

An index of association rather than of concordance was chosen because even if the attributes being measured are identical, the response scales are arbitrary and there is no reason to expect the measures to agree exactly. A correlation coefficient based on the rank rather than the raw score was chosen because of the ordinal nature of the response scales and because the relationship between variables was hypothesised to be monotonic increasing rather than linear. The hypothesis is simply that higher levels of one measure will be associated with higher levels of the other. Spearman's rank correlation coefficient fulfils these requirements.

4.2.1.3 Interpretation of the correlation coefficients

Absolute criteria for the interpretation of correlation coefficients are arbitrary since whether a correlation is trivial, moderate, or high depends as much on the underlying question as it does on the magnitude of the coefficient. However, guidelines for quantitative significance are helpful in both setting a priori hypotheses and in communicating results. In this thesis the criteria for the quantitative significance of the correlations were based loosely on the recommendations of Burnand, Kernan and Feinstein\textsuperscript{73}: $< 0.30$: insignificant; $0.30$-$0.45$: moderate; $0.45$-$0.60$: substantial; and $> 0.60$: high.
4.2.1.4 Hypotheses

*A priori* hypotheses about specific correlations were made on the basis of simple consensus among 3 of the investigators (MS, IT, PG) and are summarised in Table 4. Correlations between alternate measures of the same attribute were expected to be high, whilst measures of related but distinct domains such as fatigue and physical activity were expected to be moderate. Since the items in the PROSQOLI were selected for their pertinence to patients receiving treatment for advanced hormone-resistant prostate cancer, it was expected that each of the individual symptom and function scales should be correlated moderately with the global ratings of HRQL - the PROSQOLI LASAS for overall well-being and the EORTCI scale for global quality-of-life.
Table 4. Summary of the *a priori* hypotheses for convergent validity.

<table>
<thead>
<tr>
<th>PROSQOLI Scales</th>
<th>Validity diagonal $r \geq 0.45$ EORTCI Scales</th>
<th>Related but distinct domains</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPI (ordinal)</td>
<td>Pain</td>
<td>Fatigue</td>
</tr>
<tr>
<td>Pain (LASAS)</td>
<td>Fatigue</td>
<td>Physical Mood/Emotional</td>
</tr>
<tr>
<td>Fatigue</td>
<td>Appetite</td>
<td>Physical</td>
</tr>
<tr>
<td>Appetite</td>
<td>Constipation</td>
<td></td>
</tr>
<tr>
<td>Constipation</td>
<td>Urinary Symptoms</td>
<td>Constipation</td>
</tr>
<tr>
<td>Passing Urine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical Activity</td>
<td>Physical Function</td>
<td></td>
</tr>
<tr>
<td>Mood</td>
<td>Emotional Function</td>
<td></td>
</tr>
<tr>
<td>Family/Marriage R'ships</td>
<td>Social Function</td>
<td>Mood/Emotional</td>
</tr>
<tr>
<td>Overall Well-being</td>
<td>Global Quality of Life</td>
<td>ALL OTHERS</td>
</tr>
</tbody>
</table>


4.2.2 Multivariable relationships

4.2.2.1 Introduction and aims

The purposes of the multivariable regression were to determine the proportion of the variance in the global scales that could be explained by the specific scales, to assess the independence of the relationships between the individual specific scales and the global scales and to compare these relationships across the two instruments.

4.2.2.2 Statistical methods

The multivariable relationships between global ratings of HRQL (global scales) and ratings for the specific symptom and function domains (specific scales) were explored with multiple linear regression. For each of the 2 global scales, there were 2 analyses - one regressing the global scale (dependent variable) on the specific scales from the same instrument (predictor variables), and the other regressing the global scale (dependent variable) on the specific scales from the alternative instrument (predictor variables). For each of the 4 resulting analyses a "best" model was chosen from an exhaustive search using Mallow's Cp as the criterion to select the subset of predictor variables. Discussion of Mallows' Cp is deferred to the section on predictive validity. Linear regression assumes that the errors, but not the initial distributions, are normally distributed with zero mean. This assumption was checked by examining normal probability plots of the residuals from the final models. The relationships of interest were monotonic increasing therefore non-linear relationships were not examined. Interaction terms were neither expected nor of particular interest and were also not examined.
4.3 Findings

4.3.1 Bivariate relationships

4.3.1.1 Description of the correlation matrix

The multitrait-multimethod matrix comparing the PROSQOLI LASAS and PPI with the EORTC scales is shown in Table 5. Figure 7 is a novel graphical representation devised by the candidate, intended to facilitate the recognition of patterns, in which the size (surface area) of each circle is monotonically related to the magnitude of the correlation coefficient by the equation: surface area $\propto r^{1.2}$. This equation is based on the relationship between the actual and perceived surface area of a circle which has been estimated in several studies to be approximately: perceived = actual$^{10.8}$. MacDonald-Ross and Meihoffer have written excellent reviews of the relevant psychophysical research, including the salutary effect of providing a scale.$^{75,76}$
Table 5. The multitrait-multimethod Spearman's rank correlation matrix. The validity diagonal correlations are in bold face and the correlations between related but distinct domains are underlined.

<table>
<thead>
<tr>
<th>PROSQOLI LASAS</th>
<th>PROSQOLI PPI</th>
<th>EORTC SCALES</th>
<th>PROSQOLI PPI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pain</td>
<td>.31</td>
<td>.12</td>
<td>.21</td>
</tr>
<tr>
<td>Fatigue</td>
<td>.31</td>
<td>.33</td>
<td>.30</td>
</tr>
<tr>
<td>Appetite</td>
<td>.12</td>
<td>.33</td>
<td>.42</td>
</tr>
<tr>
<td>Constipation</td>
<td>.21</td>
<td>.30</td>
<td>.42</td>
</tr>
<tr>
<td>Passing Urine</td>
<td>.07</td>
<td>.16</td>
<td>.02</td>
</tr>
<tr>
<td>Physical Activity</td>
<td>.42</td>
<td>.36</td>
<td>.24</td>
</tr>
<tr>
<td>Mood</td>
<td>.30</td>
<td>.42</td>
<td>.43</td>
</tr>
<tr>
<td>Family/Marriage R'ships</td>
<td>.04</td>
<td>.08</td>
<td>.28</td>
</tr>
<tr>
<td>Overall Well-Being</td>
<td>.40</td>
<td>.50</td>
<td>.45</td>
</tr>
<tr>
<td>EORTC SCALES Pain</td>
<td>.59</td>
<td>.35</td>
<td>.34</td>
</tr>
<tr>
<td>Fatigue</td>
<td>.40</td>
<td>.49</td>
<td>.40</td>
</tr>
<tr>
<td>Appetite</td>
<td>.15</td>
<td>.26</td>
<td>.80</td>
</tr>
<tr>
<td>Constipation</td>
<td>.19</td>
<td>.21</td>
<td>.28</td>
</tr>
<tr>
<td>Urinary Symptoms</td>
<td>.11</td>
<td>.04</td>
<td>.04</td>
</tr>
<tr>
<td>Physical Function</td>
<td>.32</td>
<td>.42</td>
<td>.39</td>
</tr>
<tr>
<td>Emotional Function</td>
<td>.24</td>
<td>.36</td>
<td>.31</td>
</tr>
<tr>
<td>Social Function</td>
<td>.35</td>
<td>.30</td>
<td>.36</td>
</tr>
<tr>
<td>Global Quality of Life</td>
<td>.44</td>
<td>.39</td>
<td>.30</td>
</tr>
<tr>
<td>PROSQOLI PPI Pain</td>
<td>.60</td>
<td>.23</td>
<td>.22</td>
</tr>
</tbody>
</table>
Baseline Rank Correlations

PROSQOLI
- pain
- fatigue
- appetite
- constipn
- urinary
- physical
- mood
- relnships
- global

EORTCI
- pain
- fatigue
- appetite
- constipn
- urinary
- physical
- emotional
- social
- global

The size of the dot reflects the size of the correlation coefficient.
The entire matrix contains 171 correlation coefficients. The distribution of the correlation coefficients is shown in Figure 8. The median of the correlation coefficients in the matrix was 0.29 (interquartile range 0.16 to 0.41). All 8 negative coefficients involved the urinary symptom items (remembering that all scales have been transformed so that higher scores represent better quality of life - better function and less symptoms).

The magnitude of a correlation coefficient and the sample size are sufficient to calculate a p-value for the (uninteresting) null hypothesis that the true correlation coefficient is zero. The sample size is between 150 and 161 for all coefficients except those including the urinary scales. Table 6 shows p-values and approximate 95% confidence intervals for representative correlation coefficients based on a sample size of 150. The correlations for the urinary symptom scales were uniformly low (mean 0.07, range -0.17 to 0.29); the lower sample sizes for these scales does not influence the size of the correlations.
Figure 8. Histogram of the 171 correlation coefficients from the multitrait-multimethod correlation matrix.

Table 6. Confidence intervals and p-values for representative correlation coefficients from the multitrait-multimethod matrix based on a sample size of 150.

<table>
<thead>
<tr>
<th>Correlation coefficient</th>
<th>95% Confidence Interval</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>.20</td>
<td>.04 - .35</td>
<td>.01</td>
</tr>
<tr>
<td>.30</td>
<td>.15 - .44</td>
<td>.0002</td>
</tr>
<tr>
<td>.40</td>
<td>.25 - .53</td>
<td>&lt;.000001</td>
</tr>
<tr>
<td>.50</td>
<td>.37 - .61</td>
<td></td>
</tr>
<tr>
<td>.60</td>
<td>.49 - .69</td>
<td></td>
</tr>
</tbody>
</table>
4.3.1.2 Validity diagonal

The highest correlations for each scale should be those with the alternate measure of the same domain. These correlations form the validity diagonal and are shown in bold face in Table 5, and again in the first column of Table 7. The validity diagonal correlations were high for the PROSQLI LASAS for appetite, constipation, mood and pain (with the PPI); substantial for pain (with the EORTC scale), fatigue, physical activity, overall well-being and passing urine; and insignificant for family/marriage relationships. The validity diagonal included the highest correlations for the PROSQLI LASAS for pain, physical activity, appetite, constipation, mood and passing urine and for the PROSQLI PPI. The correlation between fatigue and overall well-being (.50) was slightly higher than the validity diagonal coefficient for fatigue (.49). The LASAS for overall well-being was correlated more highly with the scales for mood/emotional function and fatigue from both instruments and both the PPI and EORTC scales for pain than it was with the EORTC scale for global quality of life. The EORTC scale for global quality of life was more highly correlated with the EORTC scales for pain, fatigue, physical function, emotional function, and social function than it was with the PROSQLI LASAS for overall well-being.
Table 7. Convergent validity: important correlations from the multi-trait multi-method matrix comparing the PROSQOLI LASAS with the EORTC domains.

<table>
<thead>
<tr>
<th>PROSQOLI LASAS</th>
<th>Corresponding EORTC scales (validity diagonal)</th>
<th>PROSQOLI Overall Well-being LASAS</th>
<th>EORTC Global QL scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pain</td>
<td>.59</td>
<td>.40</td>
<td>.44</td>
</tr>
<tr>
<td>Fatigue</td>
<td>.49</td>
<td>.50</td>
<td>.39</td>
</tr>
<tr>
<td>Appetite</td>
<td>.80</td>
<td>.45</td>
<td>.30</td>
</tr>
<tr>
<td>Constipation</td>
<td>.66</td>
<td>.20</td>
<td>.27</td>
</tr>
<tr>
<td>Passing Urine</td>
<td>.44</td>
<td>.21</td>
<td>.09</td>
</tr>
<tr>
<td>Physical Activity</td>
<td>.46</td>
<td>.34</td>
<td>.31</td>
</tr>
<tr>
<td>Mood</td>
<td>.63</td>
<td>.56</td>
<td>.33</td>
</tr>
<tr>
<td>Family/Marriage R'ships</td>
<td>.11</td>
<td>.16</td>
<td>.05</td>
</tr>
<tr>
<td>Overall Well-being</td>
<td>.45</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4.3.1.3 Specific scales and global scales

Since the PROSQOLI domains were chosen on the basis of their pertinence to men with advanced hormone resistant prostate cancer we hypothesised that each of the individual symptoms and function domains would be moderately correlated with the global measures. These correlations from Table 5. and Figure 7. are shown again in Table 7. Constipation, passing urine and family/marriage relationships were the only PROSQOLI scales not correlated significantly with either of the global measures (overall well-being or global QL). The same applied for the EORTC scales for constipation and urinary symptoms; however the EORTC scale for social function was substantially correlated with both of the global measures (see Table 5. and Figure 7.).

4.3.1.4 Measures of pain

The three alternative pain scales were highly correlated (Table 8). Moderate correlations were expected between pain scales and scales for fatigue, physical activity/physical function, mood/emotional function, and global perceptions; 88% (21/24) of these correlations were ≥ 0.30. The pattern of correlations did not distinguish between the three pain scales although the highest correlations were between the EORTC pain scale and the other EORTC scales.

4.3.1.5 Related but distinct domains

Moderate correlations were expected between other (non-pain) related but distinct domains, namely: fatigue and physical activity; mood and relationships; and, passing urine and constipation. Five of the 12 predicted moderate correlations were ≥ 0.30; 6 of the 7 < 0.30 involved family/marriage relationships, passing urine and constipation (Table 9).
Table 8. Convergent validity: comparison of the three measures of pain.

<table>
<thead>
<tr>
<th></th>
<th>Spearman's rank correlation with the</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PROSQOLI Pain</td>
<td>PROSQOLI LASAS</td>
<td>EORTC Pain Likert scale</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PPI ordinal scale</td>
<td>adjectival</td>
<td>Pain</td>
<td>scale</td>
</tr>
<tr>
<td>PROSQOLI LASAS</td>
<td></td>
<td>scale</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pain</td>
<td>.60</td>
<td>.31</td>
<td>.60</td>
<td></td>
</tr>
<tr>
<td>Fatigue</td>
<td>.23</td>
<td>.32</td>
<td>.35</td>
<td></td>
</tr>
<tr>
<td>Physical Activity</td>
<td>.27</td>
<td>.43</td>
<td>.32</td>
<td></td>
</tr>
<tr>
<td>Mood</td>
<td>.30</td>
<td>.32</td>
<td>.40</td>
<td></td>
</tr>
<tr>
<td>Overall Well-being</td>
<td>.48</td>
<td>.41</td>
<td>.50</td>
<td></td>
</tr>
<tr>
<td>EORTC domains</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pain</td>
<td>.60</td>
<td></td>
<td>.60</td>
<td></td>
</tr>
<tr>
<td>Fatigue</td>
<td>.38</td>
<td>.41</td>
<td>.66</td>
<td></td>
</tr>
<tr>
<td>Physical Function</td>
<td>.36</td>
<td>.33</td>
<td>.53</td>
<td></td>
</tr>
<tr>
<td>Emotional Function</td>
<td>.32</td>
<td>.26</td>
<td>.50</td>
<td></td>
</tr>
<tr>
<td>Global QL</td>
<td>.41</td>
<td>.44</td>
<td>.65</td>
<td></td>
</tr>
</tbody>
</table>
Table 9. Spearman rank correlations between scales for related but distinct domains (not including pain).

<table>
<thead>
<tr>
<th></th>
<th>PROSQOLI Fatigue</th>
<th>PROSQOLI Physical Activity</th>
<th>EORTCI Fatigue</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROSQOLI Physical Activity</td>
<td>.36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EORTCI Fatigue</td>
<td>.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EORTCI Physical Function</td>
<td>.42</td>
<td></td>
<td>.58</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>PROSQOLI Mood</th>
<th>PROSQOLI Family/Marriage R'ships</th>
<th>EORTCI Emotional Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROSQOLI Family/Marriage R'ships</td>
<td>.29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EORTCI Emotional Function</td>
<td>.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EORTCI Social Function</td>
<td>.37</td>
<td></td>
<td>.43</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>PROSQOLI Passing Urine</th>
<th>PROSQOLI Constipation</th>
<th>EORTCI Urinary Symptoms</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROSQOLI Constipation</td>
<td>.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EORTCI Urinary Symptoms</td>
<td>.04</td>
<td></td>
<td>.04</td>
</tr>
<tr>
<td>EORTCI Constipation</td>
<td>.01</td>
<td></td>
<td>-.16</td>
</tr>
</tbody>
</table>
4.3.2 Multivariable relationships

Table 10 shows the results of the multiple linear regression analyses relating the global scales to the individual symptom and function scales. About $\frac{1}{3} - \frac{1}{2}$ of the variance in each of the global scales could be explained by the independent contributions of 3 - 4 specific symptom and/or function scales. Addition of the remaining variables to any of these "best" models did not improve their explanatory power significantly. It is unlikely that all of the residual variance of the global scales is due to error. This suggests that the global scales are measuring something additional to the specific scales and supports the use of global scales in addition to subscales.

Each global scale was more closely associated with its own specific symptom and function scales. This is probably due in part to the uniformity of the response scales within each instrument. Pain was the most consistent independent predictor of the global ratings, appearing in all 4 models, with fatigue being the next most consistent, appearing in 3 of the 4 models. The EORTC scale for fatigue did not contribute significantly ($p=0.4$) when added to the regression of the PROSQOLI LASAS for overall well-being on the EORTC scales for emotional function, physical function and pain (column 2 of Table 7).

The LASAS for overall well-being was strongly associated with both the LASAS for mood and the EORTC scale for emotional function, whereas the EORTC scale for global quality of life was not associated with either of these. The same was true of the bivariate correlations. This suggests that the PROSQOLI LASAS for overall well-being is more indicative of psychological factors than the EORTC scale for global quality of life which seems to be more indicative of physical factors.
Table 10. Multiple regression of global scales on specific scales. Each of the specific scales is significant in a model including, and therefore adjusting for, all of the other specific scales in that column. The scales have been transformed to the same metric so that the regression coefficients are comparable. The $R^2$ value indicates the proportion (range 0-1) of the variance in the global scale which is explained by the predictor variables. The adjusted $R^2$ values are all 0.01 less than the raw values: $R_{adj}^2 = 1-(n-1)\times(1-R^2)/(n-p)$.

<table>
<thead>
<tr>
<th>Specific Scales (regression coefficient)</th>
<th>PROSQOL Overall Well-being on PROSQOL specific scales</th>
<th>PROSQOL Overall Well-being on EORTC specific scales</th>
<th>EORTC Global QL on EORTC specific scales</th>
<th>EORTC Global QL on PROSQOL specific scales</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mood (.29)</td>
<td>Emotional (.36)</td>
<td>Fatigue (.26)</td>
<td>Pain (.31)</td>
<td></td>
</tr>
<tr>
<td>Fatigue (.22)</td>
<td>Physical (.30)</td>
<td>Pain (.22)</td>
<td>Fatigue (.21)</td>
<td></td>
</tr>
<tr>
<td>Pain (.22)</td>
<td>Pain (.15)</td>
<td>Physical (.15)</td>
<td>Appetite (.14)</td>
<td></td>
</tr>
<tr>
<td>Appetite (.18)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R^2$ with above domains</td>
<td>.50</td>
<td>.41</td>
<td>.54</td>
<td>.33</td>
</tr>
<tr>
<td>$R^2$ with all domains</td>
<td>.51</td>
<td>.46</td>
<td>.55</td>
<td>.35</td>
</tr>
<tr>
<td>n</td>
<td>158</td>
<td>150</td>
<td>145</td>
<td>156</td>
</tr>
</tbody>
</table>
The EORTC scale for physical function was associated with both global ratings, but the PROSQOLI LASAS for physical activity was associated with neither. Conversely, the PROSQOLI LASAS for appetite was associated with both global ratings but the EORTC scale for appetite was associated with neither. Thus, inasmuch as they are more closely related to global assessments, the EORTC physical function scale and the PROSQOLI LASAS for appetite seem superior to their counterparts.

4.4 Discussion

The bivariate rank correlations from the MTMM matrix support the convergent validity of the PROSQOLI LASAS for pain, fatigue, physical activity, appetite and mood. Although the LASAS for constipation and passing urine seem to be measuring the same attribute as their EORTC counterparts the modest associations between these scales and the global scales suggests that they are less important than anticipated. This lack of correlation for the urinary scales reflects the lack of dispersion of the scores and is not explained by the lower sample size. Possible explanations for the lack of dispersion were discussed in section 3.4.

The poor performance of the family/marriage relationships LASAS reflects the narrow range of responses and is in stark contrast to the EORTC scale for social function which performed well. The EORTC scale for social function assesses impairments of social function attributable to illness or treatment and separates family from other social contacts with two items.

The differences in the correlations between the two global scales are partly explained by differences in their wording. The PROSQOLI LASAS for overall well-being is labelled "Overall Well-being (How do you feel?)" and is anchored by "Extremely ill" and
"I feel well", whereas the EORTC global scale is the sum of separate ratings of overall physical condition and overall quality of life on a 7-point response-scale from "very poor" to "excellent". However, repeating the correlation analysis with the 2 EORTC global items separately made little difference. The two EORTC global items are more highly correlated with one another (0.8) than they are with any of the other scales. The LASAS for overall well-being is more highly correlated with the EORTC global item for overall physical condition (0.48) than it is with the EORTC item for overall quality of life (0.39), although the magnitude of this difference is modest. The EORTC scale for global quality of life was more highly correlated with several of the EORTCI specific function and symptom scales than it was with the PROSQOLI LASAS for overall wellbeing. These observations support the contention that the two global scales are not measuring the same thing but also suggest the possibility of a "method effect". A method effect occurs when the style of items or response formats in a questionnaire causes elevated associations between the attributes being measured. This might occur for the PROSQOLI if a respondent answered by marking each LASAS the same distance along the line, or for the EORTCI if they answered "quite a bit" for every 4 point Likert item. A method effect would also explain the surprising finding that the overall physical condition item of the EORTC global scale was much more highly correlated with the overall quality of life item of the EORTC global scale than it was with either the physical condition scale of the EORTCI, or the LASAS for either physical activity or overall well-being. The overall physical condition and overall quality of life items are the last 2 questions of the EORTC QLQ-C30 core questionnaire and are scored on identical numeric rating scales (from 1 to 7, anchored by very poor and excellent) which is in contrast to the 4-point Likert or dichotomous response formats of the
rest of the questionnaire.

The multitrait-multimethod matrix approach has been used widely to assess the convergent validity of health rating scales; however investigators have described the correlation matrices in a number of different ways. Most MTMM contain too many numbers to interpret at once and methods of reduction and summarisation are necessary to appreciate patterns. Bergner and Bobbitt described the results of a MTMMM assessment of the Sickness Impact Profile (SIP)\textsuperscript{14} in terms of the averages for each of the different types of correlation, for example the average of the correlations between different domains measured by the same instrument (referred to as heterotrait-homomethod by Campbell and Fiske) without showing the individual correlations. This type of summarisation is succinct and perhaps necessary for an instrument with many domains but it sacrifices detail. Ware et al and Aaronson et al have taken a more descriptive approach in reporting the convergent validity of the MOS-SF36 and EORTC QLQ-C30 respectively, showing all correlations and highlighting those which were particularly notable.\textsuperscript{17,78} This provides more information at the risk of obscuring patterns.

The approach taken in this thesis was to provide all the data to allow detailed examination with pertinent summaries to highlight important results. The graphical representation of the correlation matrix is proposed as a novel method for identifying patterns amidst a plethora of correlation coefficients.

Bergner et al used multivariable methods in assessing the SIP, claiming additional support for its convergent validity in the finding that the SIP accounted for more of the variance of measures of dysfunction than of measures of sickness (the SIP is conceptualised as a measure of the functional impact of illness rather than as a measure of the severity
of illness). In the present study, multivariable analysis was used to explore the relationships between the specific and global scales of the two instruments without specific a priori hypotheses. The multivariable analysis provides further evidence that the global scales and the specific scales are measuring related but distinct attributes. This indicates that if both specific and global attributes are of interest then both should be measured.

4.5 Conclusions

These analyses support the validity of the LASAS for pain, fatigue, appetite, mood and overall well-being. The LASAS for passing urine and constipation seem to be measuring what they purport to be measuring but may be of less importance to patients than anticipated. A substantial component of the variance in the global scales is explained by the specific symptom and function scales, particularly pain and fatigue, and to a lesser extent mood/emotional function, appetite and physical activity. The substantial variance in the global ratings which remained after inclusion of all of the specific scales supports the inclusion of a global rating scale.
5 DISCRIMINATIVE VALIDITY

5.1 Introduction and Aims

Discriminative validity refers to the ability to detect cross-sectional differences between individuals. Patients were divided into subgroups according to independent criteria and the ability of the patient-based measures to detect hypothesised differences in HRQL between these subgroups was assessed.

5.2 Methods

5.2.1 Hypotheses

A priori hypotheses were made about anticipated differences in specific domains of HRQL according to performance status, analgesic score, and serum concentrations of both hemoglobin and PSA at baseline. Predictions regarding performance status were based on previous studies assessing quality of life indices in cancer which have confirmed differences in self-rated HRQL amongst those with differing levels of performance status rated by physicians or nurses.\textsuperscript{16,17,18} Predictions for the other criteria were based on the simple consensus of 3 of the investigators all of whom are medical oncologists with experience in HRQL research (MS, PG, IT) and two of whom have specific expertise in the care of men with advanced prostate cancer (MS, IT). Specific predictions about the magnitude of the expected differences in HRQL scores were not made. The hypotheses are vague in that they reflect the expectation that the differences would be large enough to be detected with the available sample size. Comparisons of the results (p-values or differences between groups) across measures are therefore more meaningful than their absolute values.
5.2.2 Nature of the comparison

For each discriminator, the patients were divided into a high score group and a low score group; discrimination was assessed in terms of the differences between these groups in the specific HRQL scores. Differences are expressed in terms of p-values: the lower the p-value, the better the discrimination; and in terms of median differences: the bigger the difference, the better the discrimination.

5.2.3 Dichotomous discriminators

The choice of dichotomising the discriminator variables and using 2 sample comparisons was based on the aim of the analysis, the nature of the discriminators and the desire for consistency. The aim of the analysis was to detect differences between groups defined by independent criteria. The criteria, number of subjects and percentage in each group are shown in Table 11. For hemoglobin concentration the intention was to divide the patients into a symptomatic group and an asymptomatic group. The cutoff of 100 g/l was selected on the basis that it is the level beneath which transfusions are often recommended. For performance status the dichotomisation was into little or no functional impairment (grades 0-1) versus significant functional impairment (grades 2-3). There was no logical cut-point for PSA or analgesic score so they were dichotomised at their medians.

Dichotomisation of the discriminator variable discards the additional information contained in the ordinal or continuous response scales. Correlation of the variables makes use of this extra information and provides a familiar coefficient which is independent of sample size. Data from an analysis using Spearman’s rank correlation of the patient-based measures with each of the raw discriminator variables scored continuously is also presented for comparison.
Table II. Criteria, numbers of men and percentages in each group for dichotomisation of the groups according to independent clinical criteria for assessment of discriminative validity.

<table>
<thead>
<tr>
<th>Criterion scales</th>
<th>Worse Category</th>
<th></th>
<th></th>
<th>Better Category</th>
<th></th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Criterion</td>
<td>Number</td>
<td>Percent</td>
<td>Criterion</td>
<td>Number</td>
<td>Percent</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ECOG Performance Status</td>
<td>2-3</td>
<td>59</td>
<td>37</td>
<td>0-1</td>
<td>100</td>
<td>63</td>
<td></td>
<td>159</td>
</tr>
<tr>
<td>Hemoglobin g/l</td>
<td>&lt;=100</td>
<td>33</td>
<td>21</td>
<td>&gt;100</td>
<td>125</td>
<td>79</td>
<td></td>
<td>158</td>
</tr>
<tr>
<td>Analgesic Score</td>
<td>&gt;15</td>
<td>79</td>
<td>50</td>
<td>&lt;=15</td>
<td>79</td>
<td>50</td>
<td></td>
<td>158</td>
</tr>
<tr>
<td>PSA ng/ml</td>
<td>&gt;176</td>
<td>67</td>
<td>50</td>
<td>&lt;176</td>
<td>67</td>
<td>50</td>
<td></td>
<td>134</td>
</tr>
</tbody>
</table>
5.2.4 Statistical methods

Non-parametric statistics were chosen because of the ordinal nature of the scales and the expectation that the distributions would not be normal. The size of the difference between the two groups is described by the median of the differences (abbreviated as the median difference) and its non-parametric 95% confidence interval based on the Mann-Whitney statistic. The median difference is the median of all possible pairwise differences between members the two groups, not the difference between the medians of the two groups. The statistical significance of the difference between the two groups was assessed with the Mann-Whitney-Wilcoxon rank-sum test. Correlations were calculated with Spearman's rank correlation coefficient.

5.2.5 P-values

P-values were chosen as the primary summary of the differences since they reflect both the magnitude of the difference and the precision of its estimation in a simple familiar statistic. However p-values have the disadvantage of also reflecting differences in sample size. This is important when examining the p-values for the tests involving PSA for which the sample size was only 134, compared with the sample size of 158 or 159 for the other 3 discriminators. The median differences are not influenced by differences in sample size; however, their confidence intervals will be narrower for the larger samples.

5.2.6 Exploratory Analysis of Variance

Analysis of variance (ANOVA) was used to explore and describe the separate effects of the instruments, domains, and discriminators on the significance of the differences. The p-values from the primary analysis were log-transformed prior to ANOVA to correct skewness and ease interpretation. The rank correlation coefficients were not
transformed. The assumptions of the model were assessed by examination of the residuals. The aim of this analysis was purely exploratory. There were no a priori hypotheses.
5.3 Findings

The results of the assessment of discriminative validity are shown in Tables 12 and 13 which show the p-values for the PROSQOLI and EORTC instrument respectively, Figures 9-12 which show the median differences with their 95% confidence intervals, and Tables 14 and 15 which show the results of the analysis using Spearman rank correlations. Each of these will discussed in detail in the following paragraphs.

The PROSQOLI was able to detect all of the anticipated differences in HRQL in subgroups defined according to performance status, hemoglobin concentration and analgesic score (Table 12). Only 1 of the 4 anticipated differences according to PSA concentration reached borderline significance (physical activity, p = .04). The EORTC instrument yielded a similar pattern of p-values, detecting all of the anticipated differences except those according to PSA (see Table 13). The p-values for the EORTC instrument were generally lower than those for the PROSQOLI indicating better cross-sectional discrimination in this dataset.
Table 12. Discriminative Validity of the PROSQOLI. Differences in specific HRQL scores at baseline in subgroups dichotomised by independent criteria - the lower the P-value, the better the discrimination. The differences hypothesized a priori are in square brackets. Most comparisons are based on a sample size of 152-155. Comparisons involving PSA were based on a sample size of 128-130 and those involving the passing urine LASAS were based on a sample size of 71-81.

<table>
<thead>
<tr>
<th>PROSQOLI LASAS</th>
<th>P-values for differences in HRQL scores between groups dichotomised by</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prostate Specific Antigen</td>
</tr>
<tr>
<td>Pain</td>
<td>[0.5]</td>
</tr>
<tr>
<td>Fatigue</td>
<td>[.2]</td>
</tr>
<tr>
<td>Appetite</td>
<td>.14</td>
</tr>
<tr>
<td>Constipation</td>
<td>.3</td>
</tr>
<tr>
<td>Passing Urine</td>
<td>.9</td>
</tr>
<tr>
<td>Physical Activity</td>
<td>[.04]</td>
</tr>
<tr>
<td>Mood</td>
<td>.6</td>
</tr>
<tr>
<td>Family/Marriage R'ships</td>
<td>.5</td>
</tr>
<tr>
<td>Overall Well-being</td>
<td>[.14]</td>
</tr>
<tr>
<td>Geometric Mean</td>
<td>.3</td>
</tr>
</tbody>
</table>
Table 13. Discriminative Validity of the EORTCI scales. The details are as for table 12.

<table>
<thead>
<tr>
<th>EORTCI Scales</th>
<th>Prostate Specific Antigen</th>
<th>ECOG Performance Status</th>
<th>Hemoglobin</th>
<th>Analgesic Score</th>
<th>Geometric Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pain</td>
<td>.2</td>
<td>.0006</td>
<td>.01</td>
<td>.09</td>
<td>.02</td>
</tr>
<tr>
<td>Fatigue</td>
<td>.4</td>
<td>[.00003]</td>
<td>[.03]</td>
<td>.02</td>
<td>.009</td>
</tr>
<tr>
<td>Appetite</td>
<td>.7</td>
<td>.0005</td>
<td>.04</td>
<td>.006</td>
<td>.02</td>
</tr>
<tr>
<td>Constipation</td>
<td>.9</td>
<td>.04</td>
<td>.8</td>
<td>[.001]</td>
<td>.08</td>
</tr>
<tr>
<td>Urinary Symptoms</td>
<td>.16</td>
<td>.4</td>
<td>.9</td>
<td>.6</td>
<td>.5</td>
</tr>
<tr>
<td>Physical Function</td>
<td>.4</td>
<td>[.00000008]</td>
<td>[.001]</td>
<td>.02</td>
<td>.001</td>
</tr>
<tr>
<td>Emotional Function</td>
<td>.3</td>
<td>.001</td>
<td>.8</td>
<td>.04</td>
<td>.06</td>
</tr>
<tr>
<td>Social Function</td>
<td>.5</td>
<td>.01</td>
<td>.04</td>
<td>.13</td>
<td>.07</td>
</tr>
<tr>
<td>Global QL</td>
<td>.2</td>
<td>[.0001]</td>
<td>.009</td>
<td>.06</td>
<td>.07</td>
</tr>
<tr>
<td>Geometric Mean</td>
<td>.4</td>
<td>.001</td>
<td>.05</td>
<td>.03</td>
<td>.03</td>
</tr>
</tbody>
</table>
The p-values for each comparison reflect 2 aspects: the magnitude of the difference and the precision with which the difference is estimated. A non-significant p-value can reflect a small difference, an imprecise estimate of the difference (reflected in wide confidence intervals) or both. The dotplots of the median differences and their confidence intervals for the PROSQOLI LASAS reveal that the confidence intervals are quite similar across scales and discriminators (Figures 9-12). Any differences in the p-values (and therefore in discrimination) therefore reflect differences in HRQL scores between the groups rather than differences in the precision of the estimates. For example, the lack of discrimination according to PSA is due to the modest differences between the groups dichotomised according to PSA rather than imprecision due to the smaller sample size: the biggest median difference of 8 for the LASAS for physical activity compares with differences of 14 and 16 according to performance status and hemoglobin for the same scale.
Discriminative Validity
PSA <176 versus >176

Figure 9. Differences in the scores on the patient-based measures between men with blood levels of PSA < 176 and > 176. The dots indicate the median of the differences, the whiskers indicate the non-parametric 95% confidence intervals of the median difference.
**Discriminative Validity**
ECOG 0-1 versus ECOG 2-3

**PROSQOLI LASAS**
- Appetite
- Physical Activity
- Overall Well-being
- Fatigue
- Constipation
- Mood
- Pain LASAS
- Family/Marriage R'ships
- Passing Urine

**EORTC scales**
- Appetite
- Physical Function
- Global Quality of Life
- Emotional Function
- Pain Likert
- Social Function
- Fatigue
- Constipation
- Urinary Symptoms

Figure 10. Differences in the scores on the patient-based measures between men with performance status 0-1 (minimal functional impairment) versus performance status 2-3 (significant functional impairment).
Discriminative Validity
Hemoglobin <100 versus >100

PROSQOLI LASAS
Fatigue
Physical Activity
Appetite
Overall Well-being
Pain LASAS
Mood
Constipation
Family/Marriage R'ships
Passing Urine

EORTC scales
Physical Function
Social Function
Pain Likert
Fatigue
Global Quality of Life
Appetite
Urinary Symptoms
Emotional Function
Constipation

Figure 11. Differences in the scores on the patient-based measures between men with blood levels of hemoglobin < 100 and >100.
Discriminative Validity
Analgesic Score < 15 versus > 15

PROSQOLI LASAS
- Constipation
- Appetite
- Mood
- Physical Activity
- Fatigue
- Pain LASAS
- Family/Marriage R’ships
- Passing Urine
- Overall Well-being

EORTC scales
- Constipation
- Physical Function
- Fatigue
- Global Quality of Life
- Emotional Function
- Appetite
- Social Function
- Pain Likert
- Urinary Symptoms

Figure 12. Differences in the scores on the patient-based measures between men with Analgesic Scores < 15 and > 15.
Three-way analysis of variance of the log-transformed p-values was used to explore, separate and describe the effects of instrument (PROSQOLI or EORTCI), discriminative test (PSA or performance status or hemoglobin or analgesic score) and domain (pain, physical, fatigue, appetite, constipation, urinary, emotional, social and global), on the size of the p-value. Accounting for domain and discriminative test, the p-values from the EORTC instrument were about one-third the size of those from the PROSQOLI LASAS (p = .06). Accounting for domain and instrument, the differences in HRQL according to performance status were much greater than those according to the other criterion variables: the p-values for performance status were about one-tenth of those for either hemoglobin concentration or analgesic score and about one-hundredth of those for PSA (p < .0005). After adjustment for instrument and discriminative test there was still considerable heterogeneity among domains (p = .02). The largest effects (smallest p-values) were seen in physical activity/function, fatigue, appetite, and overall well-being/global quality of life. There were large differences in appetite according to performance status and analgesic score demonstrated for both instruments. Plausible differences according to analgesic score were seen with the EORTC scales, but not the PROSQOLI scales, for both physical function and fatigue. Both instruments detected differences in the global ratings according to hemoglobin concentration.
The complementary analyses using Spearman's rank correlation between the patient-based measures and the clinical criteria scored continuously are illustrated in Tables 14 and 15. The overall pattern of results is consistent with the analysis described above although the differences appear less striking, partly because of the difference in scale between correlation coefficients and p-values, and partly because the effect of smaller sample sizes for PSA is no longer apparent. Three-way ANOVA gives results almost equivalent to the analysis reported above. There is a weak tendency for the correlations to be higher for the EORTCI than the PROSQOLI scales (on average the correlation coefficients for the EORTCI are .04 higher than those for the PROSQOLI, p = .07); the strongest relationships were with ECOG performance status and the weakest were with PSA (p = .06); and, there was considerable heterogeneity among domains with the same pattern as in the analysis using p-values based on the Mann-Whitney-Wilcoxon rank-sum test.
Table 14. Discriminative Validity of the PROSQOLI scales assessed by Spearman's rank correlations with the criterion variables scored continuously. Higher correlations reflect stronger associations and better discrimination. The associations hypothesized a priori are in square brackets. All variables have been linearly transformed so that high scores are "better", thus all correlations should be positive.

<table>
<thead>
<tr>
<th>PROSQOLI LASAS</th>
<th>Prostate Specific Antigen</th>
<th>ECOG Performance Status</th>
<th>Hemoglobin</th>
<th>Analgesic Score</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pain</td>
<td>-[.01]</td>
<td>.10</td>
<td>.06</td>
<td>.08</td>
<td>.06</td>
</tr>
<tr>
<td>Appetite</td>
<td>.15</td>
<td>.25</td>
<td>.25</td>
<td>.16</td>
<td>.20</td>
</tr>
<tr>
<td>Constipation</td>
<td>-.05</td>
<td>.16</td>
<td>.08</td>
<td>[.20]</td>
<td>.10</td>
</tr>
<tr>
<td>Passing Urine</td>
<td>.01</td>
<td>-.03</td>
<td>-.11</td>
<td>-.08</td>
<td>-.05</td>
</tr>
<tr>
<td>Physical Activity</td>
<td>[.29]</td>
<td>[.29]</td>
<td>[.36]</td>
<td>.04</td>
<td>.25</td>
</tr>
<tr>
<td>Mood</td>
<td>.11</td>
<td>.24</td>
<td>.14</td>
<td>.19</td>
<td>.17</td>
</tr>
<tr>
<td>Family/Marriage R'ships</td>
<td>.13</td>
<td>.03</td>
<td>.09</td>
<td>.08</td>
<td>.08</td>
</tr>
<tr>
<td>Overall Well-being</td>
<td>[.11]</td>
<td>[.30]</td>
<td>.17</td>
<td>.02</td>
<td>.15</td>
</tr>
<tr>
<td>Mean</td>
<td>.10</td>
<td>.18</td>
<td>.15</td>
<td>.09</td>
<td>.13</td>
</tr>
</tbody>
</table>
Table 15. Discriminative Validity of the EORTCI scales assessed by Spearman's rank correlations as for Table 14.

<table>
<thead>
<tr>
<th>EORTCI Scales</th>
<th>Prostate Specific Antigen</th>
<th>ECOG Performance Status</th>
<th>Hemoglobin</th>
<th>Analgesic Score</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pain</td>
<td>[.00]</td>
<td>.32</td>
<td>.20</td>
<td>.26</td>
<td>.19</td>
</tr>
<tr>
<td>Fatigue</td>
<td>[.02]</td>
<td>[.37]</td>
<td>[.27]</td>
<td>.31</td>
<td>.23</td>
</tr>
<tr>
<td>Appetite</td>
<td>.09</td>
<td>.29</td>
<td>.29</td>
<td>.26</td>
<td>.23</td>
</tr>
<tr>
<td>Constipation</td>
<td>.03</td>
<td>.15</td>
<td>.15</td>
<td>[.30]</td>
<td>.16</td>
</tr>
<tr>
<td>Urinary Symptoms</td>
<td>-.08</td>
<td>-.01</td>
<td>-.07</td>
<td>-.07</td>
<td>-.06</td>
</tr>
<tr>
<td>Physical Function</td>
<td>[.16]</td>
<td>[.47]</td>
<td>[.41]</td>
<td>.25</td>
<td>.32</td>
</tr>
<tr>
<td>Emotional Function</td>
<td>-.04</td>
<td>.29</td>
<td>.05</td>
<td>.26</td>
<td>.14</td>
</tr>
<tr>
<td>Social Function</td>
<td>-.01</td>
<td>.26</td>
<td>.24</td>
<td>.21</td>
<td>.17</td>
</tr>
<tr>
<td>Mean</td>
<td>.01</td>
<td>.27</td>
<td>.19</td>
<td>.23</td>
<td>.17</td>
</tr>
</tbody>
</table>
5.4 Discussion

The term discriminative validity may cause confusion. Discriminant validity is a psychometric concept closely related to, and usually assessed with, convergent validity in a multitrait-multimethod correlation matrix. The discriminant validity of an index is supported when its correlations with less related indices are smaller than its correlations with more related indices. Thus discriminant validity reflects the "distinctness" of an index from conceptually distinct indices whilst convergent validity refers to the similarity of an index to conceptually similar indices. The term discriminative validity, as used in this thesis, refers to the ability of an index to discriminate between groups of individuals classified according to independent criteria. Thus the discriminative validity of an index is supported when the index is able to discern differences between groups classified according to a conceptually distinct but related measure. The psychometric synonym is known-groups validity as used by Ware et al in comparing the SF-36 scores for individuals classified according to comorbid conditions.78

The assessment of discriminative validity has been popular in the assessment of HRQL indices for cancer research. Substantial differences in HRQL between groups classified according to stage of disease and physical performance status have been found by different workers using different instruments.15-19 Aaronson et al used ANOVA to assess differences in EORTC QLQ-C30 scores according to performance status rated with the ECOG scale, weight loss and indices of treatment toxicity.17 Schipper et al used correlation coefficients to assess the relationships between the FLIC score and performance status rated with the Karnofsky scale.16 Cella et al used both ANOVA and
correlation coefficients to assess relationships between FACTS-G scores and performance status rated with the ECOG scale.\textsuperscript{18}

The choice between correlation and group comparison is arbitrary. Correlation coefficients make use of all the information in the criterion index and are optimal if the relationship is thought to be monotonic. This is the case for performance status where the underlying hypothesis is that as performance status declines, so too does quality of life, particularly physical activity, fatigue, and overall well-being. If the relationship is not monotonic then correlation coefficients will produce misleading results. In the case of hemoglobin concentration, the underlying hypothesis was that individuals with low hemoglobin concentrations would have worse fatigue and physical activity than those with normal hemoglobin concentrations. However it was not expected that individuals with higher than average hemoglobin concentrations would have better levels of physical activity or less fatigue than men with average hemoglobin concentrations. The analyses using group comparisons and correlations yielded complementary results and conclusions in the present study.

The choice of independent discriminator variables reflects the nature of the conventional clinical data assessed in this, and other, randomised trials in cancer. The measures reflect biological aspects of the disease and the host's response. The hypothesised and detected differences in HRQL reflect this biological bias. Tests of the discriminative validity of scales assessing social and emotional function depend on the availability of independent variables which are likely to be associated with differences in these attributes. Such variables were not available in the dataset for this trial.
5.5 Conclusions

The present results support the validity of the PROSQLI LASAS for fatigue, physical activity, overall well-being and constipation - all of the domains for which a priori hypotheses were made. Preferential support for these domains reflects the nature of the conventional variables against which discriminative validity was assessed. The results also suggest that the EORTCI scales provide better discrimination according to these particular clinical criteria. The consistency of the results according to PSA seen with the two HRQL instruments suggest that there is little relationship between baseline PSA and HRQL.
6 PREDICTIVE VALIDITY

6.1 Introduction

Predictive validity refers to the ability of an index to predict an important future outcome: here the degree to which baseline scores for the patient-based measures predict survival duration. An association between the index (HRQL score) and an important independent outcome (survival duration) supports the contention that the index is measuring something important: Comparability of the associations across alternative indices purporting to measure the same thing supports the contention that they are measuring the same thing.

6.2 Univariable Methods

6.2.1 Introduction and aim

The aim of the univariable analysis was to determine and compare the ability of the patient-based and conventional measures to predict survival duration. This necessitated comparison with a common metric and method: all predictor variables were dichotomised and survival distributions were compared with the logrank test.

6.2.2 Hypotheses

The a priori hypothesis was that the patient-based measures would be associated with survival duration. This was based on previous work in breast cancer, melanoma and lung cancer. Since specific predictions about the magnitude of the expected differences in survival were not made, the hypotheses are vague in that they reflect the expectation that the differences would be large enough to be detected with the available sample size. Comparisons of the results (p-values or differences in survival) across measures are therefore more meaningful than their absolute values.
6.2.3 Dichotomisation of predictor variables

For each patient-based measure the patients were divided into 2 approximately equal sized groups (roughly above and below the median): the survival durations of these groups were compared with the logrank test. Tables 16 and 17 show the criteria, numbers and percentages of men in each group for each of the patient-based scales. Performance status, analgesic score, age, time from initial diagnosis, PSA and hemoglobin concentration were treated in the same way (see Table 18). The serum concentrations of ALP, PAP, and creatinine were divided into above or below the upper limit of normal (Table 18). The categories for the blood tests (PSA, PAP, ALP, Hemoglobin and Creatinine) were based on exploratory analyses using 3 or 4 categories with amalgamation of those groups having similar survival curves; thus, the p-values for these variables should be viewed more conservatively as the categories have been selected by data-driven decisions.
Table 16. Criteria, numbers of men and percentages in each group for the dichotomisation of the PROSQOLI scales for assessment of univariable prognostic significance.

<table>
<thead>
<tr>
<th>PROSQOLI Scales</th>
<th>Worse Category</th>
<th>Better Category</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Criterion</td>
<td>Number</td>
<td>Percent</td>
</tr>
<tr>
<td>Pain LASAS</td>
<td>&lt;63</td>
<td>79</td>
<td>50</td>
</tr>
<tr>
<td>Fatigue</td>
<td>&lt;46</td>
<td>80</td>
<td>50</td>
</tr>
<tr>
<td>Appetite</td>
<td>≤75</td>
<td>80</td>
<td>50</td>
</tr>
<tr>
<td>Constipation</td>
<td>&lt;59</td>
<td>80</td>
<td>50</td>
</tr>
<tr>
<td>Passing Urine</td>
<td>&lt;89</td>
<td>43</td>
<td>50</td>
</tr>
<tr>
<td>Physical Activity</td>
<td>&lt;69</td>
<td>80</td>
<td>50</td>
</tr>
<tr>
<td>Mood</td>
<td>&lt;76</td>
<td>80</td>
<td>50</td>
</tr>
<tr>
<td>Family/Marriage R'ships</td>
<td>&lt;96</td>
<td>82</td>
<td>52</td>
</tr>
<tr>
<td>Overall Well-being</td>
<td>&lt;61</td>
<td>80</td>
<td>50</td>
</tr>
<tr>
<td>Present Pain Intensity</td>
<td>≤40</td>
<td>39</td>
<td>24</td>
</tr>
</tbody>
</table>
Table 17. Criteria, numbers of men and percentages in each group for the dichotomisation of the EORTC Scales for assessment of univariable prognostic significance.

<table>
<thead>
<tr>
<th>EORTC Scales</th>
<th>Worse Category</th>
<th>Better Category</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Criterion</td>
<td>Number</td>
<td>Percent</td>
</tr>
<tr>
<td>Pain Likert</td>
<td>≤33</td>
<td>69</td>
<td>44</td>
</tr>
<tr>
<td>Fatigue</td>
<td>≤44</td>
<td>79</td>
<td>52</td>
</tr>
<tr>
<td>Appetite</td>
<td>≤33</td>
<td>48</td>
<td>31</td>
</tr>
<tr>
<td>Constipation</td>
<td>≤33</td>
<td>58</td>
<td>36</td>
</tr>
<tr>
<td>Urinary Symptoms</td>
<td>≤67</td>
<td>42</td>
<td>51</td>
</tr>
<tr>
<td>Physical Function</td>
<td>≤40</td>
<td>67</td>
<td>43</td>
</tr>
<tr>
<td>Emotional Function</td>
<td>≤67</td>
<td>78</td>
<td>50</td>
</tr>
<tr>
<td>Social Function</td>
<td>≤50</td>
<td>70</td>
<td>45</td>
</tr>
<tr>
<td>Global Quality of Life</td>
<td>≤41</td>
<td>79</td>
<td>50</td>
</tr>
</tbody>
</table>
Table 18. Criteria, numbers of men and percentages in each group for the dichotomisation of the Traditional scales for assessment of univariable prognostic significance. uln = upper limit of normal for the laboratory.

<table>
<thead>
<tr>
<th>Traditional scales</th>
<th>Worse Category</th>
<th>Better Category</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Criterion</td>
<td>Number</td>
</tr>
<tr>
<td>ECOG Performance Status</td>
<td>2-3</td>
<td>59</td>
</tr>
<tr>
<td>Age</td>
<td>&gt;67</td>
<td>81</td>
</tr>
<tr>
<td>Analgesic Score</td>
<td>&gt;15</td>
<td>79</td>
</tr>
<tr>
<td>Years Since Diagnosis</td>
<td>≤3</td>
<td>85</td>
</tr>
<tr>
<td>Hemoglobin g/l</td>
<td>≤120</td>
<td>89</td>
</tr>
<tr>
<td>ALP uln</td>
<td>&gt;1</td>
<td>125</td>
</tr>
<tr>
<td>PSA uln</td>
<td>&gt;1</td>
<td>105</td>
</tr>
<tr>
<td>PSA ng/ml</td>
<td>&gt;176</td>
<td>67</td>
</tr>
<tr>
<td>Creatinine uln</td>
<td>&gt;1</td>
<td>25</td>
</tr>
</tbody>
</table>
6.2.4 Comparison of survival distributions

For each measure the distributions of survival duration of the two groups were compared using Kaplan-Meier survival curves and the logrank test. These methods are conventional, powerful, and assume only that the censoring is random. Their disadvantages are that they require categorisation of the predictor variables which discards information, and that they do not allow for multivariable analysis. In order to explore how this might influence the results, analyses were performed using two alternative methods - censored linear regression and Cox’s proportional hazards regression - which provide greater flexibility at the cost of additional assumptions. These issues are discussed further in subsequent sections.

6.2.5 Expression of the results

The survival durations of the two groups are illustrated with Kaplan-Meier survival curves. The strength of the association between each predictive variable and survival time is summarised by the p-value from the logrank test - the lower the p-value the stronger the relationship, and by a difference in median survivals - the bigger the difference the stronger the relationship. P-values have highly skewed distributions so they have been transformed to facilitate presentation and comparison. The transformation chosen is minus 1 times the base 10 logarithm of the p-value (-log10p). This is the same transformation that is used to express acidity in terms of pH (minus 1 times the base 10 logarithm of the hydrogen ion concentration). Higher values of -log10p reflect stronger relationships (smaller p-values). A p-value of 0.05 translates into a -log10p of 1.3. The whole number part of the -log10p value can be interpreted as the number of zeros in the raw p-value. Table 19 shows the value of -log10p for some representative values of p.
Table 19. Values of $-\log_{10} p$ for representative p-values.

<table>
<thead>
<tr>
<th>$p$</th>
<th>$-\log_{10} p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>.10</td>
<td>1.0</td>
</tr>
<tr>
<td>.05</td>
<td>1.3</td>
</tr>
<tr>
<td>.01</td>
<td>2.0</td>
</tr>
<tr>
<td>.005</td>
<td>2.3</td>
</tr>
<tr>
<td>.001</td>
<td>3.0</td>
</tr>
<tr>
<td>.0005</td>
<td>3.3</td>
</tr>
<tr>
<td>.00001</td>
<td>4.0</td>
</tr>
</tbody>
</table>
6.3 Multivariable Methods

6.3.1 Introduction and aims

The validity of an index is strengthened further if it provides predictive information above and beyond that provided by other factors. Multivariable analysis was used to address three related but distinct questions.

6.3.2 Adjusted "univariable" analyses

To address whether the patient-based measures provided additional predictive information beyond that provided by the conventional measures, each of the patient-based scales was tested individually in a multivariable model including and therefore adjusting for all of the conventional measures which were significant in the univariable analyses (p<0.05).

6.3.3 Patient-based measures alone.

To assess the independent contributions of the individual patient-based scales, "best" subsets of predictor variables were selected from starting pools including either the 9 LASAS from the PROSQOLI or the 9 analogous scales from the EORTCI.

6.3.4 Patient-based and conventional measures

To compare the predictive power of the patient-based measures with that of the conventional measures, "best" subsets of predictor variables were selected from a starting pool including all those significant in univariable analysis (p<0.05). Two analyses were performed, one including the PROSQOLI LASAS, the other including the analogous EORTCI scales. The PROSQOLI PPI and the conventional measures were included in both analyses.
6.3.5 Subset selection

An exhaustive search strategy and Mallows' Cp statistic were used to select the "best" subsets of independently significant predictor variables.\textsuperscript{83,74} The aim of these regression models is to predict duration of survival (outcome variable) from one or more predictor variables (e.g. pain score, hemoglobin &c.). Minimisation of the prediction error is therefore a reasonable basis for the selection of subsets of predictor variables. The prediction error of a model represents the difference between the actual value of the outcome variable and the value predicted by the model; the conventional definition is the sum of the squared differences between the actual and predicted values. Prediction error consists of two components: bias and random error. Addition of any variable increases random error. Addition of a variable which has predictive value reduces bias. Whether or not the addition of a particular variable improves predictions depends on the trade off between reduction in bias and increase in random error. Cp provides a numerical measure of this trade-off between bias and random error. Adequate models are those for which Cp is approximately equal to the number of parameters &/or at a minimum.\textsuperscript{70,84,85,86} Lower values indicate less bias and increased risk of overfitting, higher values indicate greater bias but less risk of overfitting.\textsuperscript{87} The subset of predictors with the minimum value of Cp for which all predictors were independently significant was selected as the "best" model: this was generally the first model for which Cp was less than the number or parameters in the model.

6.3.6 Continuous predictor variables

For these multivariable analyses the patient-based and conventional measures were treated as continuous variables. Equivalent analyses were also performed with all variables
dichotomised as for the univariate analysis.

6.3.7 Regression model

Censored linear regression was the primary method for multivariable survival analysis. The censored linear regression model was chosen in preference to Cox's proportional hazards model because of strong evidence against the proportional hazards assumption for performance status. Cox's regression model is non-parametric in the sense that it does not restrict the distributions of the survival times in a general sense; however it does assume that the hazards for different levels of predictor variables are proportional, that is, that the hazard ratios are constant over time. This is the proportional hazards assumption (PHA). Standard methods are available to test the PHA; their results are expressed by a p-value against the null hypothesis that the PHA is valid - the lower the p-value the stronger the evidence against the PHA. The p-value against the PHA was 0.0002 with performance status as a dichotomous variable and 0.005 with performance status as a continuous variable. Figure 13 shows the survival curve and log-log-survival curve for the patients dichotomised according to performance status: the variable vertical separation between the two log-log-survival curves over time reflects non-proportionality of hazards over time. Figure 14 shows the survival curve and log-log-survival curve for the patients dichotomised according to hemoglobin concentration: the constant vertical separation between the two log-log-survival curves over time reflects proportionality of hazards over time. Cox regression with stratification according to variables which violate the PHA is an alternative in this situation; however, it does not allow direct comparison of the stratification variable and other predictors which was important for our purposes.
Violation of the Proportional Hazards Assumption - ECOG PS

Hazards not proportional: vertical separation differs over time

Figure 13. Violation of the proportional hazards assumption for performance status. The upper graph shows the conventional survival curve, the lower graph the minus-log, minus-log survival curve.
Consistency with the Proportional Hazards Assumption - Hemoglobin

Hazards proportional: vertical separation is the same over time

Figure 14. Consistency with the proportional hazards assumption for hemoglobin concentration.
Censored linear regression uses maximum likelihood methods to fit a simple linear model to right-censored data based on the assumption that the errors are normally distributed with a mean of zero and common standard deviation. If a suitable transformation of the survival times can be found then the censored linear regression model has wide generality. Logarithmic and power transformations of the survival times were assessed with histograms and normal probability plots. The cube root transformation appeared suitable (see Figure 15). Other workers have described violations of the proportional hazards model and use of the censored linear regression model on cube-root transformed survival data in similar settings. The assumption of normally distributed errors with zero mean was assessed with normal probability plots of the residuals for each of the final regression models.

6.3.8 Expression of the results

The results of the multivariable analyses have been summarised by showing the variables included in the best models along with the p-values for their regression coefficients.
Figure 15. Histograms of the distributions of raw and transformed survival times.

Number of men

Number of men

survival time in years
cube root survival time in years
6.4 Findings

At the time of analysis 130 of the 161 men (81%) had died leaving 31 men with censored observations for survival time. The Kaplan-Meier estimate of the median survival for the whole group was 10.5 months with an interquartile range of 5 to 16 months. The results of the univariable analyses, summarised in Table 20, are discussed first, followed by the multivariable analyses.

6.4.1 Simple univariable associations with survival

Figure 16 shows the survival curves dichotomised according to the PROSQLI LASAS in the first column, the EORTC scales in the second column and the PROSQLI PPI scale and conventional measures in the third column. These results are summarised in Figure 17 which shows the differences between the median survivals of the dichotomised groups and Figure 18 which shows the $-\log_{10}p$ values.
Table 20. Predictive Validity: P-values for the significance of the patient-based and conventional measures as predictors of survival duration in univariable analyses.

<table>
<thead>
<tr>
<th>Patient-based Measures</th>
<th>Simple Univariable Analyses</th>
<th>&quot;Univariable&quot; Analyses Adjusted for PS, ALP &amp; Hb</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LASAS</td>
<td>EORTC</td>
</tr>
<tr>
<td>Physical Activity</td>
<td>.0003</td>
<td>.00004</td>
</tr>
<tr>
<td>Physical Function</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appetite</td>
<td>.00009</td>
<td>.0001</td>
</tr>
<tr>
<td>Overall Well-being</td>
<td>.002</td>
<td>.0003</td>
</tr>
<tr>
<td>Global QL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mood</td>
<td>.003</td>
<td>.15</td>
</tr>
<tr>
<td>Emotional Function</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fatigue</td>
<td>.004</td>
<td>.002</td>
</tr>
<tr>
<td>Family/Marriage R's</td>
<td>.006</td>
<td>.007</td>
</tr>
<tr>
<td>Social Function</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pain</td>
<td>.01</td>
<td>.008</td>
</tr>
<tr>
<td>Constipation</td>
<td>.03</td>
<td>.003</td>
</tr>
<tr>
<td>Passing Urine</td>
<td>.9</td>
<td>.8</td>
</tr>
<tr>
<td>Urinary Symptoms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Present Pain Intensity</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Conventional Measures</th>
<th>Simple Univariable Analyses</th>
<th>&quot;Univariable&quot; Analyses Adjusted for PS, ALP &amp; Hb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance Status</td>
<td>.00004</td>
<td></td>
</tr>
<tr>
<td>Hemoglobin</td>
<td></td>
<td>.0007</td>
</tr>
<tr>
<td>Alkaline Phosphatase</td>
<td>.003</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td>.05</td>
</tr>
<tr>
<td>Time from Diagnosis</td>
<td>.2</td>
<td></td>
</tr>
<tr>
<td>Analgesic Score</td>
<td>.2</td>
<td></td>
</tr>
<tr>
<td>Prostatic Acid Phosphatase</td>
<td>.4</td>
<td></td>
</tr>
<tr>
<td>Prostate Specific Antigen</td>
<td>.6</td>
<td></td>
</tr>
<tr>
<td>Creatinine</td>
<td>.9</td>
<td></td>
</tr>
</tbody>
</table>
Figure 16. Survival curves illustrating the differences in survival distributions between groups dichotomised according to the PROSQLI LASAS in column 1, the EORTCI scales in column 2, and the PROSQLI PPI and conventional measures in column3.
Predictive Validity
Actual Differences in Median Survival

PROSQOLI scales
- Physical Activity
- Pain PPI
- Appetite
- Fatigue
- Overall Well-being
- Mood
- Family/Marriage R'ships
- Pain LASAS
- Constipation
- Passing Urine

EORTC scales
- Physical Function
- Appetite
- Global Quality of Life
- Fatigue
- Pain Likert
- Constipation
- Social Function
- Emotional Function
- Urinary Symptoms

Traditional measures
- ECOG PS 0-1 v 2-3
- Hemoglobin <120 g/l>
- ALP <uln>
- Time since Dx <3 years>
- Analgesic Score <15.5>
- Age <67>
- Creatinine <uln>
- PAP <uln>
- PSA <176 ng/ml>

Figure 17. Differences in median survival between groups dichotomised according to the patient-based and conventional measures.
Predictive Validity
Logrank Test for Differences in Survival

PROSQOLI scales
- Pain PPI
- Appetite
- Physical Activity
- Overall Well-being
- Mood
- Fatigue
- Family/Marriage R'ships
- Pain LASAS
- Constipation
- Passing Urine

EORTC scales
- Physical Function
- Appetite
- Global Quality of Life
- Fatigue
- Constipation
- Social Function
- Pain Likert
- Emotional Function
- Urinary Symptoms

Traditional measures
- ECOG PS 0-1 v 2-3
- Hemoglobin <120 g/l
- ALP <uln
- Age <67
- Analgesic Score <15.5
- Time since Dx <3 years
- PAP <uln
- PSA <176 ng/ml
- Creatinine <uln

Figure 18. Differences in survival distributions between groups dichotomised according to the patient-based and conventional measures expressed as \(-\log_{10}p\) values - the higher the \(-\log_{10}p\) value the lower the p-value the bigger the difference.
There were substantial differences in survival between the high and low score groups for all of the PROSQOLI scales except the LASAS for passing urine and constipation. The PPI was the strongest predictor of the PROSQOLI scales followed by physical activity, appetite and overall well-being. The PROSQOLI LASAS for family/marriage relationships was significantly associated with survival duration despite the limited distribution of responses.

The results for the two instruments are remarkably consistent. The strongest predictors across the two instruments were physical activity/function, appetite and overall well-being/global QL. The PPI was amongst the strongest of all of the predictors, patient-based and conventional, and although the PROSQOLI LASAS for pain and the EORTCI Likert scale for pain were associated with survival, the relationships were much weaker for these two scales than for the PPI. This was reflected in the differences in both the medians and the p-values and therefore is not explained by less noise in the dichotomisation with the PPI scale. Neither of the urinary symptom scales was associated with survival. The emotional function scale of the EORTCI was only weakly associated with survival whereas the PROSQOLI LASAS for mood was quite a strong predictor. Other differences between the two HRQL instruments were minor, favouring the EORTCI for the global domain and constipation.
The PROSQOLI LASAS labelled "mood" is anchored by "very depressed" and "not depressed at all" and so is probably described more accurately as a measure of depressed mood. The EORTCI scale for emotional function consists of 4 items: "Did you feel tense?", "Did you worry?", "Did you feel irritable?", and "Did you feel depressed?". When the prognostic significance of each of these 4 items was analysed separately the association with survival duration was strongest for "Did you feel tense?" ($p=0.006$, difference in median survivals $=4.5$ months), weakest for "Did you worry?" ($p=0.4$, difference in median survivals $=3.6$ months), and intermediate for "Did you feel depressed?" ($p=0.13$, difference in median survivals $=3.5$ months) and "Did you feel irritable?" ($p=0.16$, difference in median survivals $=2.9$ months). To assess the effect of dichotomisation, censored linear regression was used to compare the EORTCI emotional function scale and the PROSQOLI LASAS scored continuously. In this analysis there was no difference in prognostic power ($p=0.008$ and 0.005 respectively). Differences between the 4 EORTCI items were less dramatic when they were scored continuously ($p=0.002$ for "Did you feel tense?" and from 0.06 to 0.10 for the other 3 items). Thus the difference between the PROSQOLI LASAS for mood and the EORTCI scale for emotional function seems more due to dichotomisation than to differences between the the scales.

Performance status, hemoglobin and ALP were the strongest predictors of survival duration amongst the conventional measures. Age was of borderline significance with younger men tending to live longer. There was no significant association between survival duration and PSA, PAP, or analgesic score.
The univariable analyses were performed in several ways to assess the dependence of the results on the details of the methods. Figure 19 shows the comparison of the results obtained from the logrank test with those obtained from censored linear regression using the same dichotomous predictor variables and expressed as $-\log_{10} p$ values. These results are further illustrated with the scatterplot of the $-\log_{10} p$ values in Figure 20. These figures show that the two methods provide very similar results although censored linear regression gives more extreme results for both performance status and the EORTCI scale for physical function. Figures 21 and 22 show the comparison of the results obtained with continuous rather than dichotomous predictor variables. Again, the findings are very consistent across methods; however there is a consistent tendency for the continuous predictors to give more significant $p$-values. Finally, in Figures 23 and 24, the results using censored linear regression are compared to those using Cox's proportional hazards model to assess univariable prognostic significance with continuous predictors. The correlation is again very high with the only notable difference being the $p$-value for the EORTCI scale for physical function which is about $10^{-9}$ with censored linear regression but about $10^{-6}$ with Cox's regression.
Predictive Validity with Dichotomous Variables

The Logrank Test v. Censored Linear Regression

PROSQOLI scales

- Pain PPI
- Appetite
- Physical Activity
- Overall Well-being
- Mood
- Fatigue
- Family/Marriage R’ships
- Pain LASAS
- Constipation
- Passing Urine

EORTC scales

- Physical Function
- Appetite
- Global Quality of Life
- Fatigue
- Constipation
- Social Function
- Pain Likert
- Emotional Function
- Urinary Symptoms

Traditional measures

- ECOG PS 0-1 v 2-3
- Hemoglobin <120 g/l>
- ALP <uln>
- Age <67>
- Analgesic Score <15.5>
- Time since Dx <3 years>
- PAP <uln>
- PSA <176 ng/ml>
- Creatinine <uln>

Figure 19. Comparison of the logrank test with censored linear regression for the assessment of the predictive significance of the patient-based and conventional measures. A \(-\log_{10} p\) value of 1.3 signifies a p-value .05. The results are highly consistent and are shown as a scatterplot in Figure 20.
Predictive Validity with Dichotomous Variables

The Logrank Test v. Censored Linear Regression

\[-\log p(\text{CLR}) = 1.1 \times -\log p(\text{logrank}) - .12\]

or \[p(\text{CLR}) = .77 \times p(\text{logrank})^{1.1}\]

Pearson's correlation = .93

Figure 20. Comparison of the logrank test with censored linear regression for the assessment of predictive significance with dichotomous predictor variables. The line 45 degree of perfect agreement is shown for comparison. The simple linear regression equation of the relationship between -log p-values for the two methods - censored linear regression (CLR) and the logrank test is also shown along with the linear correlation between the two methods. The results from the two methods are highly consistent.
Univariable Censored Linear Regression
Dichotomous Versus Continuous Predictor Variables

PROSQOLI scales
- Pain PPI
- Physical Activity
- Appetite
- Overall Well-being
- Family/Marriage R'ships
- Fatigue
- Mood
- Pain LASAS
- Constipation
- Passing Urine

EORTC scales
- Physical Function
- Appetite
- Global Quality of Life
- Fatigue
- Constipation
- Pain Likert
- Social Function
- Emotional Function
- Urinary Symptoms

Traditional measures
- ECOG PS 0-1 v 2-3
- Hemoglobin <120 g/l>
- ALP <uln>
- Age <67>
- Time since Dx <3 years>
- PSA <176 ng/ml>
- Analgesic Score <15.5>
- PAP <uln>
- Creatinine <uln>

Figure 21. Comparison of dichotomous with continuous predictor variables for the assessment of predictive significance with censored linear regression. Results with the two methods are highly consistent. The notable exceptions are the EORTCI scale for emotional function which is significant as a continuous variable but not as a dichotomous variable, and the PROSQOLI LASAS for family/marriage relationships for which the opposite is true.
Univariable Censored Linear Regression
Dichotomous Versus Continuous Predictor Variables

\[-\log p(\text{CLR continuous}) = 1.21 \times -\log p(\text{dich}) - 0.03\]

or \[p(\text{cont}) = 0.94 \times p(\text{dich})^{1.21}\]

Pearson's correlation = .90

Figure 22. Comparison of dichotomous and continuous predictor variables for the assessment of predictive significance with censored linear regression. The results from the two methods are highly consistent. On average, the analysis with continuous variables produces more significant results.
Predictive Validity with Continuous Variables
Censored Linear Regression v. Cox's Proportional Hazards Model

PROSQOLI scales

Overall Well-being
Appetite
Pain PPI
Fatigue
Physical Activity
Mood
Pain LASAS
Constipation
Family/Marriage R'ships
Passing Urine

EORTC scales

Physical Function
Global Quality of Life
Appetite
Fatigue
Pain Likert
Social Function
Emotional Function
Constipation
Urinary Symptoms

Traditional measures

ECOG PS 0-4
Hemoglobin g/l
ALP uln
Age
PAP uln
Analgesic Score
PSA ng/ml
Creatinine uln
Years Since Diagnosis

Figure 23. Comparison of censored linear regression with Cox's proportional hazards model using continuous predictor variables for the assessment of predictive significance. Results with the two methods are highly consistent.
Predictive Validity with Continuous Variables
Censored Linear Regression v. Cox's Proportional Hazards Model

\[-\log p(\text{Cox}) = 0.85 \times -\log p(\text{CLR}) + 0.31\]

or \[p(\text{Cox}) = 2.1 \times p(\text{CLR})^{0.85}\]

Pearson's correlation = 0.95

---

Figure 24. Comparison of censored linear regression with Cox's proportional hazards model using continuous predictor variables for the assessment of predictive significance. The two methods agree very closely.
6.4.2 Adjusted "univariable" analyses

The patient-based measures provided additional prognostic information beyond that provided by conventional prognostic factors (Table 20, columns 4 and 5). When added to a model including, and therefore adjusting for, all of the significant conventional factors (ECOG performance status, ALP and hemoglobin) the PPI scale, and the LASAS for overall well-being, appetite, and fatigue from the PROSQOLI each contributed significant additional prognostic information. In the equivalent analysis for the EORTCI scales physical function, global quality of life, fatigue, appetite and pain provided significant additional prognostic information.
6.4.3 Multivariable prognostic significance

The results of the multivariable analyses are summarised in Table 21. The best subset of multivariable predictors selected from the pool of PROSQOLI scales, shown in the first column Table 21, consisted of overall well-being (p = .02), appetite (p = .02) and physical activity (p = .04). The best subset of multivariable predictors selected from the pool of EORTCI scales, shown in the second column Table 21, consisted of physical function (p < .000001) and appetite (p = .01).

The best subsets selected from the pool of patient-based and conventional measures are shown in columns 3 to 6 of Table 21. The consistency of the findings for the instruments is striking. Scored as continuous variables, the best subset of predictors consisted of performance status, hemoglobin concentration, and either the PROSQOLI LASAS for overall well-being or the EORTCI scale for global quality of life (third and fourth columns of Table 21). Scored as dichotomous variables, the best subset included performance status, ALP, PPI and either the PROSQOLI LASAS for physical activity or the EORTCI scale for physical function (fifth and sixth columns of Table 21). If the PPI scale was excluded from the starting pool of dichotomous variables, then the best subset consisted of just performance status, ALP and either physical activity or physical function; that is, no other variable took the place of the PPI.
Table 21. Predictive Validity: P-values for the independent significance of the patient-based and conventional measures as predictors of survival duration in multivariable models.

<table>
<thead>
<tr>
<th>Patient-based Measures</th>
<th>selecting from patient-based measures alone</th>
<th>selecting from patient-based &amp; conventional measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>(PROSOQLI LASAS</td>
<td>EORTC)</td>
<td>continuous</td>
</tr>
<tr>
<td>Physical Activity</td>
<td>Physical Function</td>
<td>.04</td>
</tr>
<tr>
<td>Appetite</td>
<td>.02</td>
<td>.01</td>
</tr>
<tr>
<td>Overall Well-being</td>
<td>Global QL</td>
<td>.03</td>
</tr>
<tr>
<td>Present Pain Intensity</td>
<td>.005</td>
<td>.01</td>
</tr>
<tr>
<td>Conventional Measures</td>
<td></td>
<td>.005</td>
</tr>
<tr>
<td>Performance Status</td>
<td></td>
<td>.02</td>
</tr>
</tbody>
</table>
6.5 Discussion

The patient-based measures are strong, consistent predictors of survival duration and their prognostic power compares favourably with that of conventional measures. The relationship of a quality-of-life index with survival duration does not prove that the index is measuring quality-of-life. However, strong relationships between other HRQL indices and survival duration have been reported by other investigators for patients with a variety of cancers.\textsuperscript{67,68,69,70} The consistency of the results for the analogous scales across the two instruments strengthens the conclusions. This is particularly striking in the multivariable analyses where the identical domains were selected from the two instruments. The consistent ability of the patient-based measures to contribute prognostic information independent of performance status, hemoglobin and ALP is also striking. The lack of any relationship between PSA, PAP and survival duration does not bear directly on the validity of the PROSQOLI; however, it does suggest that these blood tests do not provide useful prognostic information in men with advanced hormone-resistant prostate cancer.

6.6 Conclusions

These analyses support the validity of all the PROSQOLI scales except the LASAS for passing urine and emphasise the information that patient-based measures provide above and beyond conventional measures. In addition to their primary aim as outcome measures, patient-based measures should be considered as stratification variables for clinical trials alongside conventional indices such as performance status.
CONCLUSIONS

The PROSQOLI was designed to be an outcome measure for clinical trials in symptomatic men with advanced prostate cancer, to detect and quantify the palliative effects of treatment, rather than to be a comprehensive, all-purpose measure of quality-of-life. The exclusion of role function, sexual function, and existential considerations, for example, reflects the particular treatments and questions of interest when the instrument was devised. The PROSQOLI is brief, easy to complete, and covers those aspects of HRQL felt by the investigators to be pertinent to the assessment of low-dose oral prednisone and chemotherapy with mitoxantrone.

Validity is a conditional concept - it is only meaningful in a defined context. The context includes the population, intervention and study design. Instruments are developed, validated and applied in particular contexts. The validity of an instrument cannot be judged without specifying for what and in whom it is to be used. The context may be very narrow, as is the case of the PROSQOLI, or very broad, as is the case MOS SF-36. Good validity in symptomatic men with advanced hormone-resistant prostate cancer does not guarantee good validity in men with earlier stage prostate cancer in whom, for example, pain might be less important and sexual function more important. Good validity in either or both of these populations will not guarantee good validity across the whole spectrum of prostate cancer or across the spectrum of older men in general. Even within the same population of subjects, differences between interventions, such as toxicity profiles, might influence validity. For example, nausea and vomiting, might be important in a trial of doxorubicin or cisplatin, whereas neurologic symptoms might be more important in a trial of suramin. The design of the study in which the instrument is to be used will influence
the relative importance of different aspects of validity. For example, discriminative validity is more relevant in cross-sectional studies where the aim is to distinguish between individuals at a single time point, whereas responsiveness is more relevant in longitudinal studies where the aim is detect changes over time within an individual. Thus the context in which an instrument is to be used, and the context/s in which its validity was assessed are crucial considerations which must be reexamined for each application.

In this thesis, the validity of the PROSQOLI was assessed in the context in which it was developed and applied. The study provides strong conclusions regarding the validity of the PROSQOLI in the randomised trial of mitoxantrone plus prednisone versus prednisone alone. This is the advantage of validation by application. Generalizability of the results will only become an issue when evaluating the PROSQOLI for use in subsequent studies. This will require careful consideration of the similarity between the intended context and the context in which this validation study was carried out. The results are most applicable to studies of systemic treatments without substantial toxicity applied in symptomatic men with advanced hormone-resistant disease. The degree to which these results can be generalised to other contexts will depend on the similarity of the patients, interventions, and study designs.

Validity is a relative concept - the choice of any absolute criterion for validity is arbitrary. This is particularly clear when there is no gold-standard for comparison and instruments are evaluated in terms of construct validity. Construct validity depends on assumptions which are often as tenuous as the instruments under scrutiny. What is the pass mark for such a test? How big a difference in survival should be expected in groups defined according to levels of fatigue? How big a difference in constipation should be
expected in groups defined according to analgesic intake? Answers to these questions depend in part on familiarity with the concepts and scales but any absolute criterion is arbitrary.

The validity of two instruments can be assessed and compared within the same context without having to set arbitrary absolute criteria. This framework allows statements like: "instrument A performed better than instrument B in this test of validity in that context". This kind of statement may be adequate if the question is which instrument to use rather than whether or not to use any instrument at all. Furthermore, validation, like quality of life, is multidimensional and involves multiple hypotheses and tests which must be reconciled in order to make a decision. These concepts are equally true of most scientific hypotheses. Tests of validity (observations or experiments) do not provide absolute proof or rejection. In the spirit of Hacking and Edwards, it is preferable to think in terms of a continuum of support rather than a dichotomy of acceptance or rejection.92,93

The distributions of baseline scores provide support for the use of non-parametric statistical summaries and tests. Examination of the distributions identified problems with the PROSQOLI LASAS for family/marriage relationships and passing urine. These problems were corroborated by subsequent analyses and have prompted the development of new scales for social function and urinary symptoms. There was a surprising degree of consistency in the spread of the baseline scores across domains and instruments perhaps reflecting the constrained response scales and providing a rough guide for sample size estimates.
Convergent validity was best supported for the scales assessing appetite, constipation, mood and pain. Pain, fatigue and appetite were the specific attributes most closely associated with global ratings. Psychological factors were more closely associated with the PROSQOLI LASAS for overall well-being than with the EORTC scale for global QL, reflecting the deliberate bias of the EORTC global rating towards physical condition. Although the PROSQOLI LASAS for constipation and passing urine seem to measure what they purport to, the low correlations with global ratings suggest that these domains are less important to patients. The poor performance of the PROSQOLI LASAS for family-marriage relationships in comparison with the EORTC scale for social function has prompted modifications which will be assessed in subsequent studies.

Interpretation of the multitrait-multimethod matrix was influenced by the relative homogeneity of this clinically-defined population in comparison to the relatively heterogenous demographically-defined populations for which the multitrait-multimethod matrix approach was formalised. Moderate correlations were expected and found between distinct domains; this was interpreted as confirming the association of these distinct attributes in this particular population, rather than as indicating that the scales were measuring the same (not distinct) attributes.

Tests of discriminative validity provided most support for the PROSQOLI LASAS assessing physical activity, fatigue, overall well-being, and to a lesser extent, appetite and mood. This reflects the nature of the conventional clinical criteria against which the patient-based scales were assessed. The EORTCI scales provided somewhat better discrimination than the PROSQOLI scales in general, and in particular for the scales assessing physical function and social function.
Tests of predictive validity provided strong support for the validity of all the PROSQOLI scales except that for passing urine. The strongest support was for the PROSQOLI LASAS for physical activity, appetite, overall well-being and the ordinal adjectival PPI scale for pain. The results for the EORTCI scales were highly consistent with those for the PROSQOLI; however, the EORTCI scale for physical function performed a little better, and the EORTCI scale for emotional function performed a little worse than the analogous scales from the PROSQOLI. The most striking aspect of these results was the marked superiority of the patient-based measures in comparison with the blood levels of PSA and PAP. Consistent results were obtained with a variety of methods.

The choice of the EORTC QLQ-C30 as the alternative HRQL instrument for the randomised trial was based on accumulating data on its validity in other types of cancer, its use in prostate cancer, and its adoption as the standard measure of HRQL by the Clinical Trials Group of the National Cancer Institute of Canada. Since the randomised trial was designed additional studies supporting the validity of EORTC QLQ-C30 have been published. Studies supporting the validity of other instruments, both generic and cancer specific, have also become available since the study was designed. Inclusion of other HRQL scales would have strengthened the results of this analysis. In particular, comparison against one of the widely used generic measures of health status such as the MOS SF-36 or the SIP would have provided valuable information, both in terms of internal validity and generalisability.

The randomised study provided an opportunity to compare the PROSQOLI to conventional measures. Baseline PSA levels were associated with neither HRQL nor survival duration. Comparisons between measures of HRQL and PSA were biased against
PSA by the unequal sample sizes (about 160 for HRQL versus 130 for PSA). This reflects the fact that assessment of PSA was not considered of prime importance when the randomised trial was designed. This notwithstanding, the poor performance of PSA is striking and does not lend support to its use as a surrogate endpoint for clinical trials in advanced prostate cancer.

The strong association between the domains of the PROSQOLI and survival duration add to the evidence supporting the prognostic power and importance of patient-based assessments of HRQL. Simple patient-based measures, such as the PPI or the overall well-being LASAS, should be considered as stratification variables for clinical trials in this setting, alongside conventional variables such as performance status.

The consistency of the findings for each aspect of validity using different statistical methods supports the validity of the conclusions. The consistency is partly a result of the robustness of the associations and statistical methods. It also reflects the use of statistical methods as tools to answer well-specified, a priori questions rather than using them as a means of generating questions.

The PROSQOLI is shorter, simpler, and more transparent (easier to describe and explain) than the EORTCI. Shorter instruments reduce the burden for patients, data-managers, statisticians, and the intended audience. The possible disadvantages of a shorter instrument include inadequate coverage of content and reduced reliability through lack of replication. Whether a shorter or longer instrument is preferable for a particular study depends on the particular circumstances and questions of interest. In the context of the current thesis, the questions of interest for the randomised trial were defined a priori and were addressed specifically by the PROSQOLI. Whether the PROSQOLI will be suitable
for future studies will depend on their details.

The PROSQOLI and the EORTCI also differ in terms of their time frames, the last 24-hours and the last 7 days respectively. This distinction is important in a general sense and more so in the context of rapidly fluctuating quality of life. Advanced hormone-resistant prostate cancer tends to progress relatively slowly, however, and the effects of the treatments evaluated in the randomised trial also tend to have a long time course. Shorter time frames are more indicative of states, whereas longer time frames are more indicative of traits. It is not possible to tease out the different effects of the various differences between the instruments with the data in the current study. For example, effects due to differences between Likert and linear analog scales, single versus multi-item scales and 24-hour versus 7-day time frames cannot be distinguished. In the context of the randomized trial, these distinctions were not felt to be critical and the consistency of the results despite these differences strengthened the conclusions. The same applies to the conclusions of this thesis: whether differences between the instruments are due to differences in response scales, numbers of items or time-frames can only be guessed at; the findings were remarkably consistent despite these differences.
Several important questions about the measurement properties of the PROSQOLI remain unanswered by this thesis. Some, like responsiveness, can be addressed with the data collected in the randomised trial. Others, such as content validity and reliability, cannot be addressed with the available data and represent priorities for future research.

Content validity was not addressed in the current study. Empiric data regarding content validity could have been collected as part of the randomised trial by including an instrument asking subjects to identify additional concerns not addressed by the PROSQOLI and to rate the relative importance of these additional domains and those of the PROSQOLI. Such an instrument has been designed will be administered in a subsequent study.

The high compliance rates in this study suggest that the PROSQOLI was acceptable to the men participating in the study. Personal discussions with men completing the instrument support this contention. Personal experience administering the instrument indicates that most men encountering the instrument for the first time can complete it within 5 minutes (including explanation and instruction). A formal assessment of patient acceptability and feasibility in a subset of the men in the randomised trial would have provided useful information for those concerned about the burden of formal HRQL assessment.

Reliability is an important measurement property which has not been assessed for the PROSQOLI. Test-retest reliability assessed with the intra-class correlation coefficient is the appropriate option for this single item per domain instrument; suitable data was not available in the dataset from the randomised trial. One approach to the assessment of reliability in the randomised trial would have been to administer transition scales assessing
the degree of change between successive time points for each of the domains of the PROSQOLI. Reliability could then have been assessed in those subjects indicating "no change" on the transition scales. The disadvantage of this approach is the circularity of using an unvalidated transition scale as a gold-standard for stability. A preferable approach would have been to administer the PROSQOLI on more than one occasion during a run-in period prior to starting treatment. This strategy will be incorporated in subsequent studies using the PROSQOLI.

The parallel assessment of HRQL with independent instruments in a randomised trial demonstrating an important and consistent difference provides an ideal context for the assessment of the responsiveness. This assessment has been completed and will be reported separately. In brief, the findings support the responsiveness of the LASAS for all domains except passing urine and family/marriage relationships - the same domains identified as unsatisfactorily measured in the present thesis.94

In conclusion, the present results support the cross-sectional validity of the PROSQOLI scales for pain, physical activity, fatigue, appetite, mood, overall well-being, and constipation. The cross-sectional validity of the scales for passing urine and family/marriage relationships was not supported by the data prompting the development of new scales for these domains. The patient-based measures were powerful independent predictors of survival duration, in stark contrast to PSA and PAP which have been recommended as surrogate outcome measures. This thesis adds support to the validity of both the PROSQOLI and the EORTCI as outcome measures for clinical trials in advanced prostate cancer as well as to the validity of HRQL as an outcome measure for clinical trials in general.
REFERENCES.


6. Seidman AD, Scher HI, Petrylak D, Dershaw DD, Curley T. Estramustine and vinblastine: use of prostate specific antigen as a clinical trial endpoint for hormone refractory prostatic cancer.


70. SPIDA [computer program]. Sydney, Australia: Statistical Computing Laboratory, Macquarie University; 1988-92.


Appendix 1.

**PROTOCOL CCI - NOV 22**

**Linear Analogue Scales**

Please place a vertical mark on each of the lines below, which best describes your condition **during the past 24 hours**. An example of how to complete this scale is illustrated below. Describe how you feel about freezing rain:

**FREEZING RAIN**

<table>
<thead>
<tr>
<th>Hate It</th>
<th>Love It</th>
</tr>
</thead>
</table>

1. **PAIN**  
   - Extremely  
   - Severe  

2. **PHYSICAL ACTIVITY**  
   - Completely  
   - unable to move  

3. **FATIGUE**  
   - Extremely  
   - tired  

4. **APPETITE**  
   - Not  
   - Eating  

5. **CONSTIPATION**  
   - No bowel  
   - movements  

6. **FAMILY/MARRIAGE RELATIONSHIPS**  
   - Extremely bad  

7. **MOOD**  
   - Very  
   - depressed  

8. **PASSING URINE**  
   - Causing a lot  
   - of problems  

9. **OVERALL WELL-BEING (How do you feel?)**  
   - Extremely ill  

**PRESENT PAIN INTENSITY**

Please circle the appropriate number according to how much pain you felt on average **during the past 24 hours**.

<table>
<thead>
<tr>
<th>no pain</th>
<th>mild pain</th>
<th>discomforting pain</th>
<th>distressing pain</th>
<th>horrible pain</th>
<th>excruciating pain</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>
We are interested in some things about you and your health. Please answer all the questions by circling the number that best applies to you. There are no "right" or "wrong" answers. The information that you provide will remain strictly confidential.

Please fill in your initials: .................................................................

Your birthdate (Day, Month, Year): ....................................................

Today's date (Day, Month, Year): .......................................................  

<table>
<thead>
<tr>
<th>NO</th>
<th>YES</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

1. Do you have any trouble doing strenuous activities, like carrying a heavy shopping bag or a suitcase?  
2. Do you have any trouble taking a long walk?  
3. Do you have any trouble taking a short walk outside of the house?  
4. Do you have to stay in a bed or a chair for most of the day?  
5. Do you need help with eating, dressing, washing yourself or using the toilet?  
6. Are you limited in any way in doing either your work or doing household jobs?  
7. Are you completely unable to work at a job or to do household jobs?

DURING THE PAST WEEK:

<table>
<thead>
<tr>
<th>Not at All</th>
<th>A Little</th>
<th>Quite a Bit</th>
<th>Very Much</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

8. Were you short of breath?  
9. Have you had pain?  
10. Did you need to rest?  
11. Have you had trouble sleeping?  
12. Have you felt weak?  
13. Have you lacked appetite?  
14. Have you felt nauseated?  
15. Have you vomited?  
16. Have you been constipated?  
17. Have you had diarrhea?  
18. Were you tired?  
19. Did pain interfere with your daily activities?
DURING THE PAST WEEK:

<table>
<thead>
<tr>
<th>Question</th>
<th>Not at All</th>
<th>A Little</th>
<th>Quite a Bit</th>
<th>Very Much</th>
</tr>
</thead>
<tbody>
<tr>
<td>20. Have you had difficulty in concentrating on things, like reading the newspaper or watching television?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>21. Did you feel tense?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>22. Did you worry?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>23. Did you feel irritable?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>24. Did you feel depressed?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>25. Have you had difficulty remembering things?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>26. Has your physical condition or medical treatment interfered with your family life?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>27. Has your physical condition or medical treatment interfered with your social activities?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>28. Has your physical condition or medical treatment caused you financial difficulties?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

FOR THE FOLLOWING QUESTIONS PLEASE CIRCLE THE NUMBER BETWEEN 1 AND 7 THAT BEST APPLIES TO YOU

29. How would you rate your overall physical condition during the past week?

<table>
<thead>
<tr>
<th>Rating</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very poor</td>
<td>Excellent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

30. How would you rate your overall quality of life during the past week?

<table>
<thead>
<tr>
<th>Rating</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very poor</td>
<td>Excellent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
PROSTATE MODULE

Patients sometimes report that they have the following symptoms. Please indicate the extent to which you have experienced these symptoms during the past week.

<table>
<thead>
<tr>
<th>DURING THE PAST WEEK:</th>
<th>Not at All</th>
<th>A Little</th>
<th>Quite a Bit</th>
<th>Very Much</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Have you felt pain during physical activity?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>2. Have you felt pain while sitting or lying down?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>3. Did pain wake you up at night?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4. Has pain prevented you from enjoying your family or friends?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>5. Has pain interfered with your social activities?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>6. Have you been upset by hair loss?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>7. Have you been bothered by any change in your sense of taste?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>8. Have you felt drowsy most of the time?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>9. Have you felt confused?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>10. (a) Have you taken any medication for pain?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>(b) If you took medicine for pain, did any pain remain?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>11. Have you been dissatisfied with the relief your pain medicine gave you?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>12. Have you had any difficulties urinating? (&quot;passing your water&quot;)</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>13. Has your sleep been disturbed by the need to urinate at night?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>
Comparison of the analogous HRQL scales.

**PROSQOLI** - Prostate Cancer Specific Quality of Life Instrument

**EORTC** - Core Quality of Life Questionnaire (QLQ-C30) & trial-specific module (QLM-P14)

<table>
<thead>
<tr>
<th>PROSQOLI*</th>
<th>EORTC QLQ-C30 &amp; QLM-P14</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Linear Analog Scales</strong></td>
<td><strong>Likert Scales</strong></td>
</tr>
<tr>
<td>1 Item per Domain</td>
<td>1 - 5 Items per Domain</td>
</tr>
<tr>
<td>Continuous Response (0.0-10.0)</td>
<td>4-13 Levels per Domain</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Linear analog scale label</th>
<th>HRQL domain</th>
<th>Items per Domain</th>
<th>Possible Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>physical activity</td>
<td>physical function</td>
<td>5</td>
<td>1-6</td>
</tr>
<tr>
<td>mood</td>
<td>emotional function</td>
<td>4</td>
<td>1-13</td>
</tr>
<tr>
<td>relationships</td>
<td>social function</td>
<td>2</td>
<td>1-7</td>
</tr>
<tr>
<td>pain*</td>
<td>pain</td>
<td>2</td>
<td>1-7</td>
</tr>
<tr>
<td>fatigue</td>
<td>fatigue</td>
<td>3</td>
<td>1-10</td>
</tr>
<tr>
<td>appetite</td>
<td>appetite</td>
<td>1</td>
<td>1-4</td>
</tr>
<tr>
<td>constipation</td>
<td>constipation</td>
<td>1</td>
<td>1-4</td>
</tr>
<tr>
<td>passing urine</td>
<td>urinary symptoms'</td>
<td>2</td>
<td>1-7</td>
</tr>
<tr>
<td>overall well-being</td>
<td>global qi</td>
<td>2</td>
<td>1-13</td>
</tr>
</tbody>
</table>

* The PROSQOLI also includes the adjectival Present Pain Intensity Scale, and the Analgesic Score.

' Urinary Symptoms is the only domain considered from the trial-specific module (QLM-P14).
Performance Status (ECOG/UICC)

0 Fully active, able to carry on all pre-disease performance without restriction (Karnofsky 90-100).

1 Restricted in physically strenuous activity but ambulatory and able to carry out work of a light or sedentary nature (eg, light housework, office work, Karnofsky 70-80).

2 Ambulatory and capable of all self-care but unable to carry out any work activities. Up and about more than 50% of waking hours (Karnofsky 50-60).

3 Capable of only limited self-care, confined to bed or chair more than 50% of waking hours (Karnofsky 50-60).

4 Completely disabled. Cannot carry out any self-care. Totally confined to bed or chair (Karnofsky 10-20).

ECOG = Eastern Cooperative Oncology Group,
UICC = Union Internationale Contre Cancer
Analogous grades from the Karnofsky Scale are included in brackets
CONSENT FORM

Phase 3 randomized study of mitoxantrone plus low-dose prednisone versus low-dose prednisone alone for hormone-resistant prostate cancer.

After discussion with my doctor, I understand that I have prostatic cancer which is causing symptoms and is not responsive to treatment which has reduced the stimulating effects of male hormones on my disease. Usual treatment for my cancer includes the use of medicine to control pain, low doses of prednisone (cortisone), and occasional use of radiation therapy to areas of my body that are particularly painful. Some investigators may also recommend chemotherapy, (the injection of drugs which may kill cancer cells), but this treatment has not been proven to be of benefit for my disease.

Mitoxantrone is a chemotherapy drug which we wish to test to see if it will help relieve pain and other symptoms in patients with prostate cancer. To determine the value of this treatment, I have been asked to take part in a study where I would receive, by random chance (similar to the toss of a coin), either of the following treatments:

A. Prednisone alone, taken daily by mouth.

OR

B. Prednisone and injections of mitoxantrone every 3 weeks.

I would also continue to take pain medication at a dose and frequency that is required to control my pain. As part of this study, I would complete questionnaires every 3 weeks that evaluate any pain or other symptoms that I may be experiencing. I would also undergo periodic blood tests, and if I receive mitoxantrone I will require 1 or 2 extra blood tests each month while I am taking the drug. X-rays and scans will be performed every three months as is usual for patients with this disease.

Prednisone given at low dose has few side effects. It can cause worsening of high blood pressure, ulcers or diabetes in patients with a history of these diseases, but even then can usually be given with other medication to prevent any adverse effects.

cont'd.../2
Consent Form
Page 2

Mitoxantrone is usually well-tolerated. It may cause some nausea, but this is usually mild and I will receive medicine to prevent or minimize nausea. The major side effect is a fall in the white blood count or platelets which might rarely lead to infection (for example, in my urine, throat or lung), or bleeding. This will be checked carefully and appropriate adjustments in the dose of mitoxantrone will be made. The drug may also cause some loss of hair or soreness in the mouth. Mitoxantrone can lead to heart damage in rare cases, but the drug will be stopped before sufficient dose is given to cause an appreciable risk of heart damage. The combination of mitoxantrone and prednisone causes no more side effects than mitoxantrone alone.

There are no immediate benefits to me from taking part in the study. I understand that my participation in this study is voluntary and that I am free to withdraw at any time, without prejudicing the care that I receive. I also understand that I will not be identified in any publication or presentation that results from this study.

An independent physician with whom I may discuss this study is

Dr. _______________________ telephone # _______________________

I agree to take part in this study.

Signature: ______________________  Date: ______________________

Witness: ______________________