COMPARING THE RESPONSIVENESS OF TWO HEALTH-RELATED QUALITY OF LIFE INSTRUMENTS IN A PHASE III RANDOMIZED CLINICAL TRIAL OF MEN WITH PROSTATE CANCER (NCIC CTG PR.3): THE EORTC QLQ-C30+3 WITH PR17 TRIAL SPECIFIC CHECKLIST VERSUS THE FACT-P

By

Yvonne Murray

A thesis submitted to the Graduate Program in Epidemiology

In conformity with the requirements for

the degree of Master of Science

Queen's University

Kingston, Ontario, Canada

(April, 2015)

Copyright © Yvonne Murray, 2015

Abstract

Background: The EORTC QLQ-C30 and the FACT are two commonly used Health-Related Quality of Life (HRQL) instruments in cancer clinical trials, but there is limited data comparing them. The NCIC CTG PR.3 clinical trial compared Androgen Deprivation Therapy (ADT) alone with ADT plus radiation therapy (ADT + RT) in prostate cancer patients. In a PR.3 sub-study, we conducted a comparison of the EORTC QLQ-C30+3 and prostate module (PR17) to FACT-P by employing a cluster randomization of 29 participating North American centers to HRQL instrument used on the PR.3 clinical trial.

Purpose: To compare the responsiveness of two HRQL instruments to short-term radiation effects and long-term hormone effects in men treated for locally advanced prostate cancer on a clinical trial.

Methods: 311 patients randomized to the PR.3 sub-study were included for analysis. HRQL was assessed at baseline, 6 monthly (for 2 years), then annually; compliance exceeded 85% to three years. The ability of each HRQL instrument to detect RT toxicity was determined by comparing mean change scores (ADT vs. ADT + RT arms) at 6 months by HRQL instrument (Wilcoxon rank-sum). The ability of each instrument to detect proportions changed (at 6 or 36 months) was determined by calculating proportions (clinically meaningful change defined as 10% change from baseline) then comparing between instrument groups (chi-square). Finally, we compared instruments on time to clinically meaningful worsening of HRQL using Kaplan-Meier survival curves/Cox regression.

Results: The FACT-P detected significant between-treatment arm differences in urinary symptom change scores at 6 months. The EORTC QLQ-C30+3/PR17 detected significant between-treatment arm differences in diarrhea and bowel/rectum symptom changes at 6 months. For functional domains and fatigue, no significant between-instrument differences were observed in proportions of patients

improved/stable and worsened at and up to 36 months. However, the FACT-P reported a faster rate of clinically meaningful HRQL decline for physical and role/functional domains.

Conclusions: When randomly assigned to patients participating in a clinical trial, the FACT-P and EORTC QLQ-C30+3/PR17 instruments differed in responsiveness to changes in urinary and bowel symptoms attributable to radiotherapy. The FACT-P was more responsive to change in physical and role function over time.

Co-Authorship

This thesis is the work of Yvonne Murray under the supervision of Dr. Harriet Richardson and Dr. Michael Brundage.

Acknowledgements

I would like to thank a number of individuals, for the support they have provided me while completing the requirements of this degree.

To my thesis supervisors Dr. Harriet Richardson and Dr. Michael Brundage, I could not have asked for better supervisors. You always made time in your busy schedules for me, and provided me with valuable guidance and advice throughout this process. You have always instilled confidence in me and provided me with useful feedback in a very positive way, and for that I am extremely grateful. I have learned a great deal from both of you. To Dr. Bingshu Chen, thank you for your statistical guidance and your support of this project.

To the staff and faculty of the Department of Public Health and to Queen's University School of Graduate Studies, thank you for your support and patience with me as I completed this degree.

Finally, a special thanks to my family. To my wonderful husband Adam for your support and understanding throughout all the years of my part-time graduate studies. You are amazing. To my 3-year-old son Xavier, for being patient all the times mommy had to work. And to my parents, in-laws, sisters and brothers, thank you for your support, encouragement and help with Xavier throughout this process. I would not have been able complete this degree without all of you.

Table of Contents

List of Figures

Chapter 3 Methods

Chapter 4 Results

List of Tables

Chapter 3 Methods

Chapter 4 Results

List of Abbreviations

ADT	Androgen Deprivation Therapy
CMC	Clinically Meaningful Change
EORTC	European Organization for Research and Treatment of Cancer
EORTC QLQ-C30	European Organization for Research and Treatment of Cancer Quality of Life Questionnaire Core 30
EORTC QLQ-C30+3	European Organization for Research and Treatment of Cancer Quality of Life Questionnaire Core 30 +3
EORTC QLQ-PR25	European Organization for Research and Treatment of Cancer - Prostate specific module
ES	Effect Size
FACT	Functional Assessment of Cancer Therapy
FACT-G	Functional Assessment of Cancer Therapy - General measure
FACT-P	Functional Assessment of Cancer Therapy - General measure with prostate cancer subscale
HR	Hazard ratio
HRQL	Health-Related Quality of Life
LHRH	Luteinizing hormone-releasing hormone
MID	Minimum Important Difference
NCIC CTG	NCIC Clinical Trials Group
PCS	Prostate Cancer Subscale
PR17	Prostate trial specific checklist
PRO	Patient-reported outcome
PSA	Prostate Specific Antigen
QOL	Quality of Life
RT	Radiotherapy
SD	Standard deviation
SF-36	Short-Form (36) Health Survey
SRM	Standardized Response Mean
SSQ	Subjective Significance Questionnaire

Chapter 1

Introduction

1.1 Overview and rationale

Health-related Quality of Life (HRQL) has become an important outcome in cancer clinical trials. HRQL is a multidimensional construct that makes up a personal perception of well-being and functioning (physical, psychological, cognitive and social) as affected by wellness, illness, treatment, ability, infirmary, quality of, and satisfaction with, care (1). An unique aspect of HRQL is that it is a patient reported outcome (PRO), that is, an assessment of an aspect of a patient's health status that comes directly from the patient (i.e., without the interpretation of the patient's responses by a physician or anyone else) (2) and as such is typically measured by patients completing brief questionnaires (HRQL instruments). HRQL has the advantage to potentially better describe and quantify treatment benefits and toxicities (compared to conventional reporting of toxicity by clinical and/or research staff), as only PROs can provide the appropriate understanding of the impact of toxicity on patients' roles, functioning and degree of 'bother'. HRQL has been shown to have significant added value in clinical trials research, in terms of choosing the 'best' treatment, enriching the understanding of patient experiences and improving clinical trial methods (1). HRQL is now commonly integrated in phase III clinical trials, often as a secondary endpoint.

HRQL instruments undergo extensive construction and validation processes before they are considered to be valid and credible to the scientific community. There are a number of HRQL instruments that have been developed and validated for use in cancer research. The EORTC QLQ-C30 (European Organization for Research and Treatment of Cancer Core Questionnaire) and the FACT (Functional Assessment for Cancer Therapy) are two of the most popular instruments used in cancer trials and have both been extensively validated internationally (3).

There is limited data available, however, with respect to comparing the responsiveness of these two instruments and no consensus as to which one is superior. In fact, there is no 'gold standard' for HRQL measurement in cancer patients at this time.

In 1995, the NCIC Clinical Trial Group (NCIC CTG) launched the PR.3 study, a randomized controlled trial that aimed to compare Androgen Deprivation Therapy (ADT) alone to ADT with radical radiation therapy (RT) in men with locally advanced prostate cancer. While overall survival was the primary endpoint of this study, health-related quality of life (HRQL) was determined to be an important secondary endpoint. At the time of the PR.3 study conception there was lack of consensus as to the most appropriate instrument to use to measure HRQL. Thus a sub-study was included in the design to address the question of the relative responsiveness of two of the most commonly used instruments in cancer clinical trials. This sub-study included North American patients only, and randomization by study centre was used to allocate instruments. North American participating centres were randomized to administer one of two instruments to their patients, either the EORTC QLQ-C30+3 with a prostate trial specific checklist (PR17) or the FACT-P (FACT general measure with prostate-specific subscale). Both of these instruments had been well validated, however thus far there have been no clinical trials comparing the responsiveness of these two HRQL instruments in patients with prostate cancer. The randomized design for allocating instruments in the PR.3 study provided a unique opportunity to compare the two instruments with respect to their performance in this population of prostate cancer patients. This thesis aimed to compare the 2 instruments using descriptive and statistical comparisons in this randomized subset of patients.

1.2 Thesis objectives

1. To describe the HRQL findings by instrument allocation (EORTC QLQ-C30+3 with PR.17 and FACT-P) and treatment arm (ADT and ADT + RT) as follows:

a. Describe mean HRQL scores at baseline, 6 months (short term) and 36 months (long term)
b. Describe the amount of missing HRQL data at baseline, 6 months and 36 months

2. To determine if there are differences in the ability to detect the incremental impact of RT on HRQL between instruments:
    a. By comparing the differences in mean change scores from baseline between the ADT and the ADT+RT arm for each instrument at 6 months, for HRQL symptom domains/items relevant to short term effects of RT
    b. By comparing the differences in proportions of patients worsened, improved and stable from baseline between the ADT and ADT+RT arm for each instrument at 6 months, for HRQL symptom domains/items relevant to short term effects of RT
3. To determine if there are differences in the ability to detect change in HRQL over time between instruments:
    a. By comparing the proportion of patients improved, stable and worsened from baseline in each instrument group at 6 months, for HRQL symptom domains/items relevant to the short term effects RT:
        - With treatment arms (ADT and ADT + RT) pooled
        - In the ADT + RT arm only
    b. By comparing the proportion of patients improved/stable and worsened from baseline in each instrument group for HRQL functional domains and fatigue (relevant to long term ADT effects):
        - At the 36 month assessment
        - Cumulatively, up to and including the 36 month assessment
    c. By comparing the time to first clinically meaningful HRQL decline between instruments groups for HRQL functional domains and fatigue (relevant to long term ADT effects)

1.3 Overview of study design

This thesis project utilized data collected for the PR.3 Study, a multi-centre non-blinded randomized trial conducted by the NCIC Clinical Trials Group. The purpose of the PR.3 study was to evaluate any possible benefit from the addition of external beam radiation therapy to the treatment of patients with locally advanced cancer of the prostate who had not had a radical

prostatectomy and are receiving hormonal therapy. The primary outcome of the study was overall survival and secondary objectives included disease specific survival, time to disease progression, symptomatic local control, and HRQL. PR.3 participants were randomized to receive hormonal treatment (Androgen Deprivation Therapy or ADT) or hormonal treatment plus radiation therapy (ADT +RT). In the HRQL sub-study, the North American centres that participated in the PR.3 study were randomized to administer one of two instruments to their patients, either the EORTC QLQ-C30+3 with a 17-item prostate trial specific checklist (PR17) or the FACT-P (FACT with prostate-specific module). HRQL instruments were administered to patients at baseline (prior to randomization), every 6 months for the first 2 years, then annually thereafter. The patients included in the analyses for this thesis project were eligible patients randomized from North America and that completed the HRQL instrument that their site was allocated to at baseline.

1.4 Thesis organization

This thesis uses the traditional style, and has five chapters. Subsequent to this introduction, chapter two follows which is a summary of the literature surrounding prostate cancer treatment, health-related quality of life, the EORTC QLQ-C30 and FACT instruments, and measuring responsiveness in HRQL. Chapter three describes the methods used for this project, including study design, scoring of questionnaires and statistical analyses. Chapter four outlines the findings of the project and, finally, chapter five is a discussion of the results in the context of relevant literature, as well as strengths, limitations and conclusions. It also includes 6 appendices as follows: Appendix A- the EORTC QLQ-C30+3 (26) and PR17 questionnaires, Appendix B - the FACT-P questionnaire (28), Appendix C – Study Power, Appendix D – Construction of domain/item scores for the EORTC QLQ-C30+3/PR17 vs. the FACT-P, Appendix E - Symptom domain construction for the EORTC QLQ-PR25 vs the PR17 Trial Specific Checklist and Appendix F – Ethics approval.

Chapter 2

Literature Review

2.1 Locally advanced prostate cancer treatment

Prostate cancer is the most common cancer among Canadian men (excluding non-melanoma skin cancers) and the 3rd leading cause of death from cancer in men in Canada (4). It is estimated that in 2014 23,600 men will be diagnosed with prostate cancer and 4000 men will die from this disease in Canada (4). There is no universally accepted definition of locally advanced prostate cancer; the term is loosely used to encompass a spectrum of disease profiles that show high-risk features (5). A commonly used definition is 'clinical evidence of tumour extension beyond the prostatic capsule with no obvious involvement of other organs, nodal disease, and/or distant metastases' (6). Although there has been a decline in the number of men presenting with locally advanced disease due to prostate-specific antigen (PSA) screening, up to one-third of newly diagnosed cases are considered to be locally advanced at the time of diagnosis (5, 7, 8). Compared to those with localized prostate cancer, men with locally advanced disease generally have a worse prognosis. The 15-year survival rate has been observed to be 81% in men with localized disease at diagnosis, versus 57% in those with locally advanced disease (9).

The optimal management of locally advanced prostate cancer still remains unclear, despite many advances in research over the last few decades. The use of androgen deprivation therapy (ADT), also referred to as hormonal therapy, is well established however the utility of radical radiation therapy (RT) in combination with ADT has been historically controversial. Androgen deprivation therapy in prostate cancer includes orchiectomy (surgical castration), luteinizing hormone-releasing hormone (LHRH) analogs, LHRH agonists and anti-androgens. The goal of ADT is to reduce levels of androgens (male hormones) in the body, (or to block their effectiveness in the case of anti-androgens) preventing them from reaching prostate cancer cells

where they promote growth. Clinical trial results emerged in the 1990s which suggested that ADT combined with RT provides additional benefit when compared to RT alone (10, 11) thus giving impetus to the use of combined therapy. However, these results were unable to address the exact role of RT as they did not include an ADT only arm. Another randomized trial attempted to address the role of RT by randomizing patients to orchiectomy alone, RT alone or combined orchiectomy and RT, and although no differences in survival were found between the 3 treatment groups (12), the study was not completed due to poor accrual and the number of patients was not sufficient to detect clinically relevant survival differences. Thus there was a still a need to determine if RT combined with ADT provides benefit over ADT alone. As RT is also associated with additional side effects, there was also a need to weigh these against the possibility of any benefit.

To address these questions, a randomized controlled trial was conducted by the NCIC Clinical Trial Group (NCIC CTG) comparing ADT alone to ADT with RT in men with locally advanced prostate cancer (PR.3). This trial accrued 1205 patients from 1995- 2005. The results of the 2nd interim analysis of this study showed that that the addition of RT to ADT significantly improves survival at 7 years (HR=0.77, p=0.033) (13). This was consistent with another study of very similar design, which showed a prostate specific and overall survival benefit with the addition of RT to endocrine therapy (14). While overall survival was the primary endpoint of the PR.3 study, health-related quality of life (HRQL) was determined to be an important secondary endpoint.

2.2 Health-related quality of life

Health-related Qualify of life (HRQL) is a multidimensional construct that makes up a personal perception of well-being and functioning (physical, psychological, cognitive and social) as affected by wellness, illness, treatment, ability, infirmary, quality of, and satisfaction with, care (1). It is the area of Quality of Life (QOL) research that focuses on health-related parameters and

is led by medical outcomes researchers, as opposed to other areas of QOL that focus on broader determinants of well-being, such as sociological and economic determinants, and are led by medical sociologists (15). HRQL is one type of outcome referred to as patient reported outcomes (PROs), which are assessments of any aspect of a patient's health status that comes directly from the patient (i.e., without the interpretation of the patient's responses by a physician or anyone else) (2). As such, HRQL is usually measured by patients completing brief questionnaires (HRQL instruments) about their quality of life. HRQL researchers have used a variety of conceptual HRQL models to guide their research (16, 17, 18). Wilson and Cleary's model of HRQL is the most cited conceptual framework in the HRQL literature (16) and combines two paradigms; biomedical and social sciences (17). Ferrans, Zerwic, Wilbur and Larson (18) published a revision of the Wilson and Cleary model that retained five major domains: biological, symptoms, function, general health and overall HRQL and explicitly clarified the definitions for individual and environmental characteristics that are associated with HRQL outcomes along the casual pathway. These conceptual models facilitate the understanding of the associations between clinical variables and HRQL outcomes, and thus are useful for guiding HRQL research and selection of appropriate measurement instruments for a given component of the model (18).

HRQL research has made much progress in the last few decades. Schwartz et al. identified 4 main areas where HRQL research has contributed to quality cancer care: to assess treatment outcome and qualify survival, to assess late problems, to predict mortality and to support transfer of information, all of which has led to HRQL research becoming increasingly accepted by the medical community and society at large (15). Specifically, HRQL has been shown to have significant added value in clinical trials research, as illustrated by Au et al. in a review of NCIC CTG Phase III studies. In this review, HRQL was found to have added value in terms of choosing the 'best' treatment, enriching the understanding of patient experiences and improving clinical

trial methods (1) . This review highlighted trials where HRQL outcomes have supported the primary outcome of the study as well as those that have counterbalanced the primary outcome, in both cases helping patients and clinicians decide on the most appropriate therapy. It is also thought that HRQL may better describe and quantify treatment benefits and toxicities (compared to conventional reporting of toxicity by clinical and/or research staff), as only PROs can provide the appropriate understanding of the impact of toxicity on patients' roles, functioning and degree of 'bother'. This is evidenced in a number of studies where the patients reported more symptoms, or greater severity of symptoms, than did clinicians assessing the same patients (1). For these reasons HRQL is now commonly integrated in phase III clinical trials, and the NCIC CTG includes HRQL in almost all Phase III studies they conduct.

2.3 HRQL in men with prostate cancer and the impact of ADT and RT

HRQL can be influenced by numerous factors, such as morbidity and treatment (19). It is well documented that prostate cancer treatment has a significant impact on HRQL. ADT is associated with a number of adverse effects including sexual dysfunction, fatigue, anemia, loss of bone density, muscle atrophy and alterations in myocyte contraction. Cross-sectional and longitudinal studies have consistently reported a decrease in self-reported physical function with ADT use (20). In a prospective matched cohort study, Alibhai et al. further added to this knowledge, showing that ADT resulted in worse objective physical function (using standard physical performance measures) as well as worse self-reported physical function scores (as measured with the SF-36 HRQL instrument) compared to control groups (20). RT is also associated with a number of side effects which include sexual dysfunction, bowel and bladder irritation, and obstructive symptoms such as frequency, urgency and incontinence (21). RT has been shown to negatively impact disease-specific domains of HRQL, namely sexual dysfunction, bowel dysfunction and urinary function scores (22).

2.4 HRQL instruments

2.4.1 Development and validation

For HRQL to be useful it must be measured with appropriate instruments and methods (1). HRQL Instruments used in clinical practice and research have three basic purposes: to discriminate among individuals along a continuum of health, illness, or disability; to predict outcome or prognosis; and to evaluate within person change over time (23). HRQL instruments are considered *discriminative* when they used to measure cross-sectional differences in quality of life between patients at a point in time, and *evaluative* when their purpose is to measure longitudinal changes in HRQL within patients over an extended period of time (19). The approach to construction and testing of an HRQL instrument is very much dictated by the purpose, and thus an instrument developed for one purpose may not be as useful for other purposes. The process of constructing and validating an HRQL instrument is quite extensive, and involves first selecting the item pool (items or questions to be included), reducing the number of items, choosing response options, and then determining the reproducibility/reliability, validity and responsiveness (23). These scientifically rigorous approaches must be applied to questionnaire development, selection and administration for the results to be considered valid and credible to government agencies and the scientific community.

Reliability, for a discriminative instrument, is a way of quantitating the signal-to-noise ratio. A discriminative instrument may be deemed reliable if the variability in scores between 2 groups of patients (the signal) is much greater than the variability within patients (the noise) (19). For evaluative instruments on the other hand, the signal–to-noise ratio is an indication of the instruments responsiveness (ability to detect change). The signal in this case is the magnitude of the difference in score in patients who have improved or deteriorated and the noise is the extent to which patients who have not changed provide more or less the same scores (19). Validity of an HRQL instrument (the ability to measure what it is intended to measure) can be determined by

evaluating the extent to which results correspond to a gold standard, when a gold standard exists. This is referred to as criterion validity. As no gold standard for HRQL exists, HRQL researchers tend to rely on construct validity, which is the ability of an instrument to measure the intended construct and is the most rigorous approach to establishing validity when criterion-based measures are not possible. A construct is a theoretically derived notion of the domain(s) we want to measure (19). For a discriminate instrument, construct validity may be established if the instrument is able to distinguish between 2 groups of patients receiving different treatments, for example, which are known to differ in toxicity. For an evaluative instrument, construct validity is shown when changes in the instrument correlate with changes in other related measures in the theoretically derived predicted direction and magnitude (19).

HRQL instruments can be generic (meaning they are designed to be used across the full range of medical conditions, populations, or interventions) or disease specific (for examining a particular condition or disease state, such as cancer) (24). Disease specific measures may be particularly useful in measuring the responsiveness of patients to interventions in a clinical practice setting (24). HRQL instruments are questionnaires usually made up of a number of items or questions, which add up to form a number of domains or dimensions. A domain or dimension refers to the area of behaviour or experience that we are trying to measure (19) such as physical, social, psychological and cognitive. Instruments may also include a 'global' rating of HRQL consisting of one or more questions pertaining to overall quality of life, as well as symptom specific items/measures, which may also consist of one or more questions.

There are a number of HRQL instruments that have been developed and validated for use in cancer research. The EORTC QLQ-C30 (European Organization for Research and Treatment of Cancer Core Questionnaire) and the FACT (Functional Assessment for Cancer Therapy) are two of the most popular instruments used in cancer patients and have both been extensively validated internationally (3). In 1988, Neil Aaronson proposed the "core plus module" approach for

cancer-specific quality of life, the idea being that a global quality of life index would be administered with a smaller additional disease specific module (specific to prostate cancer, for example). This allowed for a consistent overall quality of life assessment to be made, while at the same time dissecting the dysfunctions identified with measures specific to each disease (25). The European Organization for Research and Treatment of Cancer (EORTC) established this approach, whereby the 'core' questionnaire (the EORTC QLQ-C30), which contains questions relevant to the functional HRQL of cancer patients, is supplemented with a 'module', which contains questions specific for disease site, treatment modality or QOL dimension. The FACT follows a similar approach, where a disease specific subscale is added to the core FACT-G (FACT general measure) to form a disease-specific questionnaire (the FACT-P, for example, for prostate cancer).

2.4.2 The EORTC QLQ-C30 instrument

In 1986, the European Organization for Research and Treatment of Cancer (EORTC) initiated a research program to develop an integrated, modular approach for evaluating the quality of life of patients participating in international clinical trials (26). By 1987, the first generation of the EORTC quality of life questionnaire was developed (the EORTC QLQ-C36). This instrument was designed to be cancer specific, multidimensional in structure, appropriate for self-administration, applicable across a range of cultural settings and suitable for use with additional site or treatment-specific modules (27). Further development of this questionnaire resulted in the second generation core questionnaire, the EORTC QLQ-C30 (version 1) , which consisted of 30 items that assess five functional dimensions (physical, role, cognitive, emotional and social), three symptom dimensions (fatigue, pain, nausea/vomiting) and global quality of life, as well as a number of single items assessing additional symptoms commonly reported by cancer patients (dyspnoea, loss of appetite, insomnia, constipation and diarrhea) . By 1993 the EORTC had completed the validation of this instrument, which involved administering the questionnaire to

305 patients with non-resectable lung cancer in 13 countries, both before treatment and once during treatment, to assess the instrument's reliability and validity. The data from this study supported the hypothesized scale structure, meeting the minimal standards for reliability (Cronbach's alpha coefficient >0.7). Validity was demonstrated by a moderate inter-scale correlation (indicating that the scales were assessing distinct components of the QOL construct), by the ability of the functional and symptom measures to discriminate between patients with different clinical status and by showing statistically significant changes in physical and role functioning, global quality of life, fatigue, and nausea and vomiting, for patients with worsening or improving performance status during treatment (26). The current version of the EORTC QLQ-C30 is version 3.0, however only minor changes have been made since version 1.0.

2.4.3 The FACT-P instrument

The Functional Assessment of Cancer Therapy (FACT) measurement system, in parallel to the development of the EORTC QLQ-C30, began development in 1987. This system began with the creation of a generic CORE questionnaire called the Functional Assessment of Cancer Therapy-General (FACT-G), a 28 item questionnaire which assesses five domains (physical, social, emotional and functional well-being and relationship with doctor). From 1987 to 1992, the development and validation took place in 5 phases – item generation, item review/reduction, scale construction/piloting, initial evaluation and additional evaluation. The FACT-G was shown to be valid (as it showed high correlation with similar measures completed at the same time, and it was able to differentiate between patients according to stage of disease), reliable (as evidenced by high test-retest correlation coefficients) and also sensitive to change (as it was able to distinguish between patients that had declined performance status rating (PSR), improved PSR and stable PSR over time) (28). A prostate cancer subscale (PCS) was developed and tested in 1997, consisting of 12 items that ask about symptoms and problems specific to prostate cancer. These questions were added to the FACT-G thereby comprising the FACT-P. Validity was tested

in 2 independent samples of prostate cancer patients, with different disease stages. The first sample contained 34 patients with advanced hormone refractory prostate cancer who were enrolled in an investigational study using two oral agents in the treatment of their progressive disease, and the second sample included 86 patients who were seeking a second opinion about their disease (20% had evidence of metastasis or recurrence, 78% had localized disease, and 2% had presented with prostate problems without having had a positive biopsy result). The PCS was shown to be internally consistent (acceptable alpha coefficient) indicating that items in a given scale are measuring the same dimension. Validity was confirmed by the ability to discriminate patients by disease stage, performance status and baseline prostate specific antigen (PSA) level and sensitivity to change in performance status and PSA scores was demonstrated (29). Of note, validity sample 1 (which included 34 patients being treated for progressive disease) was used to demonstrate sensitivity to change in performance status rating and PSA level. This supported the use of the FACT-P in HRQL evaluation for men undergoing therapy for prostate cancer. The current version of the FACT-P is version 4.0 which does not differ significantly from the FACT-P earlier versions.

2.5 Clinically meaningful change and responsiveness of HRQL instruments

Responsiveness or 'sensitivity to change' is an essential property of a measuring instrument defined as the ability to detect a clinically meaningful change, such as a change that clinicians or patients think is discernible and important (30). An instrument's responsiveness is considered to be an important aspect of validity (31) and is a key criterion when choosing an HRQL instrument (32). In a clinical trial, the impact of treatment on HRQL is typically assessed by determining extent of the change from baseline, and comparing treatment groups with respect to this change. However a statistically significant difference, in HRQL measurement, may or may not equate to a clinically meaningful difference (33). Therefore determination of clinically meaningful difference in HRQL assessments is critical for clinicians to determine the impact of treatment on

patient's HRQL. Interpretation of HRQL results has been historically problematic, and the concept of clinically meaningful change (CMC) in HRQL has been the focus of much research over the last 2 decades (34, 35). Osoba et al. advises that it should be decided, *a priori*, the magnitude of change (cut-off point) that will be considered to be a clinically meaningful change (CMC) in HRQL scores and the duration that this change will need to persist in order to consider the HRQL response as being 'improved', 'worsened' and 'stable' (35). There is a great deal of variation in terms of the methods that have been used to determine what constitutes a clinically meaningful change and the responsiveness of an instrument (30, 33, 36). Most methods involve either an anchor-based or a distribution-based approach.

Anchor-based methods compare score changes with established external standards termed 'anchors'. For example, in one study designed to determine the significance to breast and small cell lung cancer patients of changes in HRQL scores (using the EORTC QLQ-C30), a subjective significance questionnaire (SSQ) was completed in which the patients rated their perception of change since the last time they completed the EORTC QLQ-C30. For each category of change in the SSQ, the corresponding differences were calculated in the EORTC QLQ-C30 mean scores and effect sizes were determined. This study found that a mean change score of 5-10 (out of 100, on the EORTC QLQ-C30) was associated with a 'a little' change as perceived by patients and as reported on the SSQ, 10-20 with a 'moderate' change, and greater than 20 'very much' change (33). This provided some guidance to researchers as to how much change on the HRQL assessment corresponds to a perceivable change to the patient, and provided a basis for calculating the sample sizes required to detect specified changes in clinical trials.

Distribution-based methods evaluate the dispersion of scores in the target population as an estimate of the scale's inherent variability, thus offering a likely range for the minimum important difference (MID) (36). Distribution-based methods interpret results in terms of the relation between the magnitude of effect and some measure or measures of variability in the results (37).

The magnitude of effect may be the difference in an individual or group's score before or after treatment or the difference in score between treatment and control groups. Guyatt et al. in a review, concluded that neither anchor-based or distribution based approaches are perfect but that the use of multiple strategies is likely to enhance the interpretability of a given instrument (37). However, these different approaches (with many different instruments and among many different cancer types) have yielded strikingly similar answers, being that a change from 5% to 10% of the scale breadth (or in general, 0.5 of a standard deviation) is perceptible to patients as a meaningful change (33, 35, 37, 38, 39, 40). Norman et al., in a systematic review of the literature identifying studies that computed a minimally important difference (MID), found that out of 38 studies all but 6 studies showed that the MID estimates were close to one half a standard deviation (SD) (40). The researchers in this article conclude that the value of 0.5 SD can therefore serve as a default value for important patient-perceived change on HRQL measures used with patients with chronic diseases (40).

MIDs have been determined specifically for certain scales within some HRQL instruments, including the FACT-P and EORTC QLQ-C30. MIDs for scores generated by the FACT-P were determined using data from a phase III trial of metastatic prostate cancer patients using both distribution-based and anchor-based methods (36). MIDs in this study were estimated for FACT-P total score, Trial Outcome index (TOI) score, prostate cancer subscale (PCS) score, pain-related score and FACT Advanced Prostate Symptom Index (FAPSI), but not for any of the functional domain (physical, functional, social, emotional) scores. Maringwa et al. (2011) conducted a study to determine minimal clinically important differences (MCID) for some EORTC QLQ-C30 scales in brain cancer patients using an anchor-based approach, and provided different estimates for improvement and deterioration. The scales analysed included three of the functional domains, and the MCID estimates were as follows (for improvement and deterioration respectively, on a 100-point scale): physical (6, 9), role (14, 12) and cognitive (8,8) (41). These results were in line with

the 5-10% range considered to be clinically significant in other studies (33, 39, 42). Another study by Bedard et al. aimed to determine the minimum important difference (MID) for the EORTC QLQ-C30 in patients with advanced cancer, using both anchor and distribution-based methods, and generated MID estimates for physical functioning (7.2), role functioning (13.5) and cognitive functioning (9.1) which were fairly consistent with the Maringwa et al study (43).

2.6 Comparisons of the EORTC QLQ-C30 and FACT instruments

The EORTC QLQ-C30 and the FACT have four functional subscales in common. Although the labels of these subscales do not match exactly and the construction of the common domains differ, they are considered to be comparable in the sense that they are intended to measure the same dimension. The common functional subscales between the two instruments include physical functioning (QLQ-C30)/physical well-being (FACT), social functioning (QLQ-C30)/social/family well-being (FACT), emotional functioning (QLQ-C30)/ emotional well-being (FACT) and role functioning (QLQ-C30)/ functional well-being (FACT) (3). These 2 instruments have been compared in a number of studies, in an attempt to provide guidance to clinical researchers in choosing between the two. Luckett et al. conducted a systematic review identifying all articles reporting on psychometric properties and information to assist interpretability as well as collated information on content, scale structure, accessibility and availability for both instruments. Comparisons of reliability, validity and responsiveness were undertaken to inform recommendations. One questionnaire was not recommended over the other, however there were important differences noted which may inform choice for a particular study (44). A number of studies identified in this review had compared the 2 instruments by correlation or distribution scores (3, 45, 46, 47, 48, 49, 50). The largest differences noted were in the measurement of social HRQL, as low correlations were observed between the EORTC QLQ C-30's social functioning and FACT-G's social functioning scores (<0.30). For the other functional domains, correlations were found to be moderate (0.30 – 0.49) to high (≥0.5). One of

these studies, by Holzner et al. (2001), examined the EORTC QLQ C-30 and FACT-G for comparability using four groups of cancer patients. Only low to moderate intercorrelations between the common subscales of the 2 instruments were found, and in some disease groups, there were actually substantial differences in the common subscales (47). Thus, there is some evidence that differences exist between the 2 instruments in terms of the results they provide.

Minimal literature is available on the responsiveness of the EORTC QLQ-C30 and the FACT-G. Osoba et al. (1994), in a study evaluating the psychometric properties and responsiveness of the EORTC QLQ-C30 in patients with breast, ovarian and lung cancer, concluded that the instrument was responsive to severity of disease (evidenced by the differences in scores between those patients with and without distant metastases) and to chemotherapy (evidenced by a decrease in physical, role, social and global functions as well as nausea/vomiting and fatigue after chemotherapy) (51). This study however evaluated the EORTC QLQ-C30 only, and did not compare responsiveness with the FACT instrument.

A few studies have compared the responsiveness of both instruments. Conroy et al. attempted to assess the responsiveness of individual subscales within the EORTC QLQ-C30 and FACT-H&N (the FACT-G plus Head and Neck Cancer subscale). 87 patients were given the FACT-H&N and the EORTC QLQ-C30 during both the first and the last week of radiotherapy. Responsiveness to change was analysed using variations in scores and the standardized response mean (SRM). Their findings indicated that the most responsive subscales were the FACT-G Physical and Functional Well-Being and the EORTC QLQ-C30 global score, because they showed the highest relative variation of scores between the end and the beginning of radiotherapy (52). Another study by Uwer et al. compared the responsiveness of the EORTC QLQ-C30 to the FACT-C (core FACT-G with colorectal cancer subscale) in 121 patients with colorectal cancer, 71 of which were undergoing chemotherapy and 56 of which were undergoing radiation therapy. Patients were administered both questionnaires before the first, third and fourth courses of

chemotherapy (for the patients receiving chemotherapy) or before the first fraction and during the last week of radiotherapy (for patients receiving radiotherapy). In this study, the patient's overall assessment of his/her change in health status was the reference criterion to evaluate responsiveness of the instrument. Responsiveness statistics included the standardized response mean (SRM) and the effect size (ES), calculated for those patients with improved health. The standardized response mean was calculated as the mean change in scores between baseline and follow-up divided by the standard deviation (SD) of this change (53). The findings indicated that the EORTC QLQ-C30 physical, role, emotional and cognitive domains and the FACT-P functional well-being domain were responsive to change in the chemotherapy group, as indicated by SRM and ES indicators greater than 0.5 (reflecting moderate ability to detect an effect of chemotherapy treatment) (30). However in the radiotherapy group neither the EORTC QLQ-C30 nor the FACT-C were shown to be responsive (30) thus indicating that responsiveness of an instrument may vary by the type treatment. Neither of these studies (30, 52) however, directly compared the 2 instruments statistically in terms of their responsiveness.

Recognizing that there was a lack of head-to-head comparisons, King et al. (2014) recently compared the responsiveness of the EORTC QLQ-C30 with the FACT-G in a secondary analysis using data from a trial which randomized patients with mixed cancer diagnoses at variable stages of disease and treatment to Medical Qigong (breathing and movement exercises) or usual care (control group) to evaluate the impact of Medical Qigong on quality of life. In this study, a responsiveness index (RI) was calculated for each instrument by dividing the mean change in the intervention arm by the standard deviation of the change in the control arm. The mean difference in RI between instruments was then calculated for the comparable EORTC QLQ-C30 and FACT-G scales. There was a statistically significant difference observed between the 2 instruments only for the social domain, for which the EORTC QLQ-C30 was shown to be more responsive compared to the FACT-G. The FACT-G however had a higher RI (indicating greater

responsiveness) for the physical, functional/role and global scores, though the differences were not statistically significant. For the emotional domain, the EORTC QLQ-C30 was more responsive than the FACT-G but this difference was also not statistically significant (54).

2.7 Summary and rationale

The literature suggests that there may be some differences between the EORTC QLQ-C30 and FACT instruments in terms of the results that they provide. These 2 instruments are considered to be the most commonly used instruments to measure cancer-specific HRQL, and are regarded as comparable with respect to 4 main functional domains they both measure. However to date there is no consensus as to which instrument performs better, and no 'gold standard' for HRQL measurement in cancer patients. A limited number of studies have directly compared the 2 instruments in terms of their relative responsiveness (or sensitivity to change). Among those that have compared responsiveness, the methods vary and the results are somewhat inconsistent. Furthermore, data has suggested that responsiveness may differ depending on the study population and treatment type. Thus, there is a need for more research comparing responsiveness of these 2 instruments in different cancer populations and with different types of cancer treatments. There are no studies to date that have addressed the relative responsiveness of 2 HRQL instruments in patients with prostate cancer who are receiving hormonal therapy (with or without radiation therapy). This project will directly compare these two widely used HRQL instruments for cancer, using descriptive and statistical methods. This study is unique from earlier studies in that randomization (by centre) was used to allocate instruments, thereby allowing a direct comparison in two similar groups of patients in a clinical trial. The methods used to assess responsiveness in this study are also unique. The results from this project therefore will provide additional insight with respect to the similarities and differences between the 2 instruments and contribute to the state of knowledge in this area.

Chapter 3

Methods

3.1 Study purpose and objectives

The purpose of this study is to compare the responsiveness of two HRQL instruments previously utilized in a randomized controlled trial of men with locally advanced prostate cancer, in a subset of patients randomly allocated (by study centre) to an HRQL instrument. Specifically, the objectives are:

1. To describe the HRQL findings by instrument allocation (EORTC QLQ-C30+3 with PR17 and FACT-P) and treatment arm (ADT and ADT + RT) as follows:
    a. Describe mean HRQL scores at baseline, 6 months (short term) and 36 months (long term)
    b. Describe the amount of missing HRQL data at baseline, 6 months and 36 months
2. To determine if there are differences in the ability to detect the incremental impact of RT on HRQL between instruments:
    a. By comparing the differences in mean change scores from baseline between the ADT and the ADT+RT arm for each instrument at 6 months, for HRQL symptom domains/items relevant to short term effects of RT
    b. By comparing the differences in proportions of patients worsened, improved and stable from baseline between the ADT and ADT+RT arm for each instrument at 6 months, for HRQL symptom domains/items relevant to short term effects of RT
3. To determine if there are differences in the ability to detect change in HRQL over time between instruments:
    a. By comparing the proportion of patients improved, stable and worsened from baseline in each instrument group at 6 months, for HRQL symptom domains/items relevant to short term effects of RT:
        - With treatment arms (ADT and ADT +RT) pooled
        - In the ADT + RT arm only

b. By comparing the proportion of patients improved/stable and worsened from baseline in each instrument group for HRQL functional domains and fatigue (relevant to long term ADT effects):
    - At the 36 month assessment
    - Cumulatively, up to and including the 36 month assessment
c. By comparing the time to first clinically meaningful HRQL decline between instruments groups for HRQL functional domains and fatigue (relevant to long term ADT effects)

3.2 Study design

3.2.1 NCIC CTG PR.3 study

This thesis project utilized data collected for the PR.3 study, a multi-centre non-blinded randomized trial conducted by the NCIC Clinical Trials Group. 1205 patients were enrolled to the PR.3 study from the United Kingdom (n=844) and North America (n=361) from 1995 to 2005. The purpose of the PR.3 study was to evaluate any possible benefit from the addition of external beam radiation therapy to the treatment of patients with locally advanced cancer of the prostate who had not had a radical prostatectomy and are receiving hormonal therapy. The primary outcome of the study was overall survival and secondary objectives included disease specific survival, time to disease progression, symptomatic local control, and HRQL.

PR.3 participants were randomized to receive hormonal treatment, referred to as androgen deprivation therapy (ADT) alone or ADT plus radiation therapy (ADT +RT). ADT included either a bilateral orchiectomy or an LHRH agonist. An oral anti-androgen was required to be taken for a minimum of two weeks if an LHRH agonist was given. Patients randomized to RT were to be given 65-69 Gy over 35-37 days. Stratification factors for the PR.3 study included study centre, initial PSA level, choice of hormonal therapy (orchiectomy or LHRH agonist), method of node staging (clinical or radiological or surgical), Gleason score, and prior hormone

therapy (excluding orchiectomy). HRQL instruments were administered to patients at baseline (prior to randomization), every 6 months for the first 2 years, then annually thereafter until death.

3.2.2 Study population

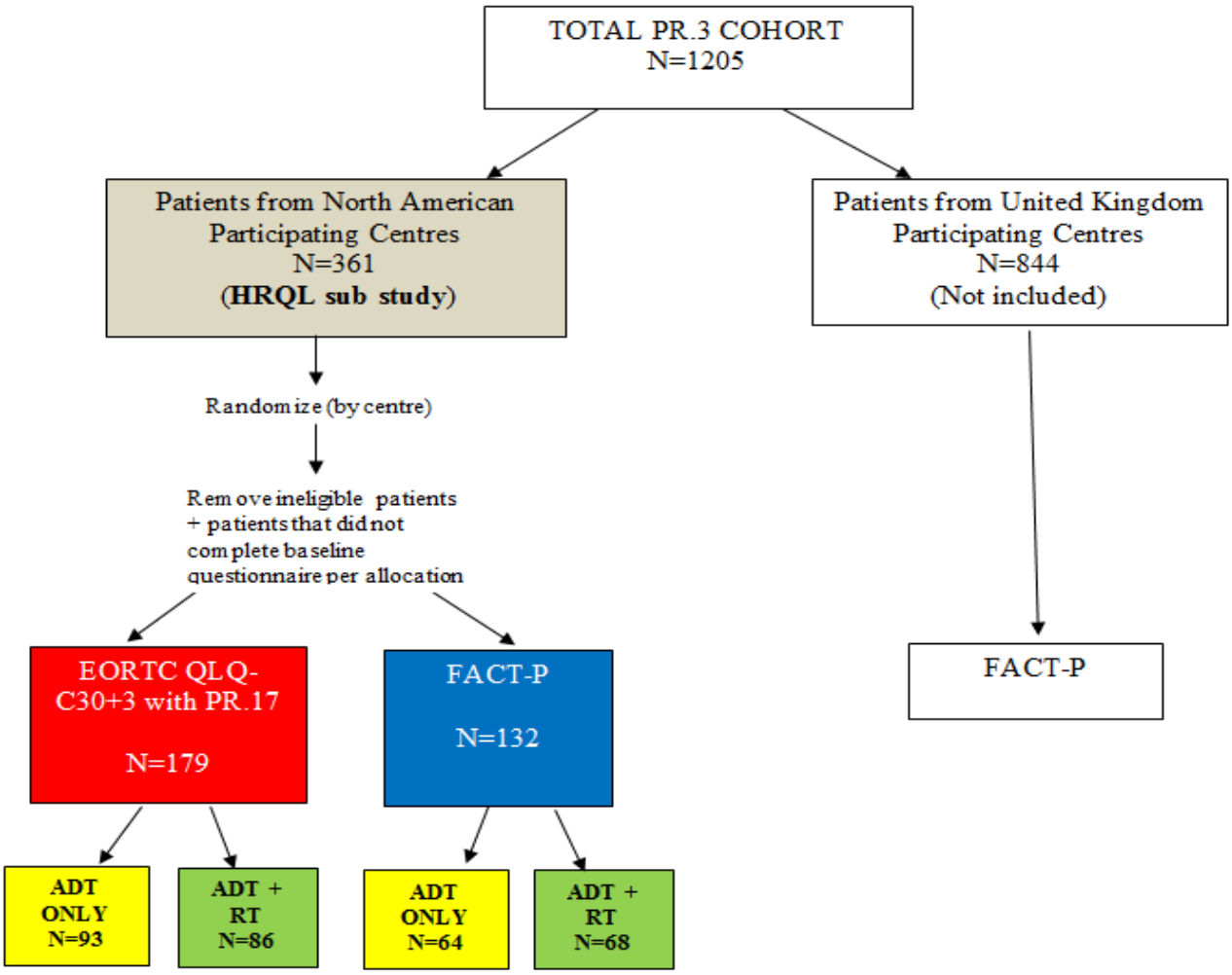
This study population for this thesis project is comprised of a subset of the study participants from the PR.3 study. Participants were eligible for the PR.3 study if they had locally advanced adenocarcinoma of the prostate (stage T3 or T4, N0 or NX, M0 or T2 PSA >40 ug/L or T2 PSA>20 ug/L and Gleason>=8) and had not had a radical prostatectomy. Participants must also have completed the pre-randomization QOL questionnaire (FACT-P or EORTC QLQ-C30+3 with PR17) and must have been willing to complete future questionnaires (unless illiterate in English or French).

In terms of the HRQL instruments, it was decided at the beginning of the PR.3 study that all of the sites participating from the UK would use the FACT-P while participating North American centres would be randomized to use one of two instruments for administration to all patients at their centre – either the EORTC QLQ-C30+3 with a 17 item trial specific checklist (PR17) or the FACT-P. Centres were stratified by the expected accrual size before randomization to HRQL instrument. This randomized subset of participants enrolled from North American make up the population for this study, as shown in Figure 3.1.

3.2.3 Exclusion criteria for thesis project

Patients were excluded from the study sample if they did not complete a baseline HRQL assessment, if they did not complete the correct HRQL instrument (i.e. the instrument their centre was allocated to), or if they were determined to be ineligible for the main PR.3 study.

Figure 3.1: PR.3 HRQL sub-study design and study population



3.3 Variables

The data necessary for the analyses in this study was extracted from the main PR.3 database. The main variables extracted included the following: Patient ID number, Patient Treatment Allocation (ADT or ADT + RT), Eligibility status (Y or N), and patient HRQL data (including time of assessment and patient responses to each question). A variable was created to designate the questionnaire completed (EORTC QLQ-C30+3 with PR.17 or FACT-P). Potential covariates also extracted were baseline characteristics including age at allocation, PSA level, method of

lymph node staging, Gleason score, clinical stage, prior hormonal therapy (Y or N), choice of hormonal therapy (on study), rectal exam normal (Y or N) and ECOG performance status.

3.3.1 Exposure variable

The main exposure variable in this study is the HRQL instrument the participant completed as per the site's allocation, either the EORTC QLQ-C30+3 with PR17 or the FACT-P.

The EORTC QLQ-C30+3 is the 'version 2' 33-item core questionnaire, which arose following international testing of the EORTC QLQ-C30 (version 1) when refinement of the questionnaire was recommended by adding three new test items. From the core EORTC QLQ-C30+3, 5 functional domain scores and 3 symptom domain scores are generated as well as global health status/QOL. The PR17 was a 17 item trial specific checklist designed specifically for the PR.3 study to be completed with the EORTC QLQ-C30+3. The PR.17 contains 2 symptom domains, 5 single items and 2 conditional items specific to prostate. At the time of the study initiation, there was no validated prostate specific module available for use with the EORTC QLQ-C30 thus the creation of this checklist (there now exists a module that has been developed and validated by the EORTC for use with the QLQ-C30, named the QLQ-PR25).

The FACT-P version 2 was used in this study, which consists of the 33-item core FACT-G plus a 12 item prostate specific subscale (PCS). The core FACT-G generates scores for 5 functional domains as well as a total FACT-G score. From the PCS, a prostate cancer subscale score can be generated, as well prostate specific symptom scores and a number of single items. The FACT-P also provides a Trial Outcome Index (TOI) score and a total FACT-P score.

Subjects were administered the instruments (either the FACT-P or the EORTC-QLQ C30+3 with PR17) by study staff and were asked to complete it independently at each visit, before being seen by the health care provider.

3.3.2 Outcome variable

The objective of this study is to compare the performance of the 2 HRQL instruments by evaluating the HRQL scores. Thus, the HRQL scores represent the outcome. The full complement of functional domain/symptom domain scores were not evaluated in this project, but rather those scales that were thought to best allow a comparison of responsiveness between to the 2 instruments. The functional domain and symptom domain/item scores chosen for evaluation were those that were common to both instruments, those expected to be impacted by RT in the short term, and those expected to be impacted by ADT in the long term. This decision was based on previous knowledge of the effects of RT and ADT on certain aspects of HRQL and also informed by the results of the main PR.3 study HRQL analysis. The following scores were selected for inclusion in this analysis:

EORTC-QLQ C-30+3 and PR.17**:**

Functional Scales: physical functioning, role functioning, emotional functioning, social functioning

Symptom scales: fatigue, urine symptoms, bowel/rectum symptoms

Single item symptoms: constipation, diarrhea, urination at night

FACT-P

Functional scales: physical well-being, functional well-being, emotional well-being, social/family well-being

Symptom scales: urine problems

Single item symptoms: bowel trouble, fatigue

3.3.3 Scoring of instruments

EORTC QLQ-C30+3 and PR.17

The EORTC QLQ-C30+3/PR17 questionnaires are included in Appendix A. For the first 7 items of the QLQ-C30+3 the respondent either indicates 'yes' or 'no'. For the following 22 items the respondent indicates a response on a scale of 1 to 4 (1=not at all, 2=a little, 3=quite a bit, 4=very much), based on their experience in the past week. The PR17 follows the same format as the majority of the EORTC QLQ-C30, where the patient indicates a response on a scale of 1 to 4 (1=not at all, 2=a little, 3=quite a bit, 4=very much) based on their experience in the past week. Table 3.1 below indicates the question numbers used to generate each score, the component of the questionnaire from which the score is generated, and the algorithm used to calculate the score (which is expressed as a value out of 100).

Table 3.1: Scoring of domains/items on the EORTC QLQ-C30+3/ PR.17

	Domain/ Item	Questionnaire component	Question Numbers	Algorithm for Score
Functional Domains	Physical functioning	EORTC QLQ-C30+3	1, 2, 3, 4, 5, 6, 7	Score=100-(((Total score for the answered questions/(Total questions answered))-1)*100)
	Role functioning	EORTC QLQ-C30+3	26, 27	Score=100-(((Total score for the answered questions/(Total questions answered))-1)*100/3)
	Emotional functioning	EORTC QLQ-C30+3	21, 22, 23, 24	Score=100-(((Total score for the answered questions/(Total questions answered))-1)*100/3)
	Social functioning	EORTC QLQ-C30+3	28, 29	Score=100-(((Total score for the answered questions/(Total questions answered))-1)*100/3)
Symptom Domains	Fatigue	EORTC QLQ-C30+3	10, 12, 18	Score=(((Total for the answered questions/(Total questions answered))-1)*100/3)
	Urine symptoms	PR.17	34 to 41	Score = ((Total for the answered questions/(Total questions answered))-1)*100/3
	Bowel/rectum symptoms	PR.17	42, 43	Score = ((Total for the answered questions/(Total questions answered))-1)*100/3
Single Items	Constipation	EORTC QLQ-C30+3	16	Score= (Answer to the question-1)*100/3
	Diarrhea	EORTC QLQ-C30+3	17	Score= (Answer to the question-1)*100/3
	Urination at night	PR.17	49	Score=(Answer to the question-1)*100/3

FACT-P

The FACT-P questionnaire is included in Appendix B. For each item in the FACT-P, the respondent provides a response on a scale of 0 to 4 (0=not at all, 1=a little bit, 2=somewhat, 3=quite a bit, 4=very much) all which represent the past week. The first 33 questions make up the FACT general measure (FACT-G) and questions 34 to 46 comprise the Prostate Cancer Sub-scale, although collectively the questionnaire is referred to as the FACT-P. Similar to the EORTC QLQ-C30+3, scores for the FACT P are generated from 2 or more questions for functional/symptom domains and from 1 question for single items and expressed as a value out of 100, according to the algorithms in Table 3.2 below.

Table 3.2: Scoring of domains/items on the FACT-P

	Domain/ Item	Questionnaire component	Question Numbers	Algorithm for Score
Functional Domains	Physical well-being	FACT-G	1 to 7	Score=100-((Total score for the answered questions/(Total questions answered))*100/4)
	Functional well-being	FACT-G	26 to 32*	Score=100-((Total score for the answered questions/(Total questions answered))*100/4)
	Emotional well-being	FACT-G	20 to 24**	Score=100-((Total score for the answered questions/(Total questions answered))*100/4)
	Social/family well-being	FACT-G	9 to 15***	Score=100-((Total score for the answered questions/(Total questions answered))*100/4)
Symptom Domains	Urine problems	PCS	42, 43, 44	Score = ((Total for the answered questions/(Total questions answered))*100/4)
Single Items	Bowel trouble	PCS	41	Score = (Answer to the question)*100/4
	Fatigue	FACT-G	1	Score = (Answer to the question)*100/4

For both the EORCT-QLQ C30+3/PR.17 and the FACT-P, domain scores were considered missing if more than half of the items used to generate domain score are unanswered. For single item symptoms, the score was considered missing if the item is unanswered.

3.4 Defining clinically meaningful change

A clinically meaningful change was defined conservatively as a change of ≥10% of the scale breadth (32), therefore a change of 10 points (as all scores are within a range of 0 to 100). For functional scales, high scores represent higher functioning (better HRQL) and for symptom scales, high scores represent lower functioning (worse HRQL). Therefore, for functional scales, patients were classified with improved HRQL if they reported an increase of ≥10 points from baseline. Conversely, patients were classified with worsened HRQL if the reported score represented a decrease of at least 10 points from the baseline score. Patients that did not have an increase or decrease of at least 10 points from baseline were considered stable. For symptoms scales/items, an increase of ≥10 points from baseline classified the patient as 'worsened' and a decrease of ≥10 points classified the patient as 'improved'. A sensitivity analysis (for objectives 3B and 3C), in which the clinically meaningful cut-point was reduced from 10 points to 7 points, was done to test the robustness of the results.

3.5 Statistical analysis

Statistical analysis was carried out using SAS 9.3 and SAS 9.4.

3.5.1 Objective 1: Mean HRQL scores and missing data

From the raw HRQL data (consisting of the patient responses to each item on the questionnaire, at each assessment), scores were first generated for the selected domains/items per the scoring methods referenced in section 3.3.3, for each participant and at each assessment time point. The mean and corresponding standard deviation was then computed for each domain/item score at baseline, 6 months, and 36 months by treatment allocation and by instrument allocation.

Missing data was evaluated at each of these time points, for each instrument. Data could be considered 'missing' in the either of the following scenarios:

a) The entire questionnaire was not completed by the patient at a protocol mandated assessment time (i.e. *all* HRQL data would be missing for that assessment). This was evaluated by computing the compliance at each assessment, by instrument group. Compliance is calculated by dividing the number of patients who completed a HRQL assessment at the designated time point by the total number of patients 'expected' to complete it at that time point (i.e. the number of patients still on study at the time of the scheduled assessment) and expressed as a percentage.

b) The questionnaire was partially completed but too few items within the questionnaire were answered such that a domain/item score could not be generated (i.e. the domain/item score would be considered missing). Domain scores were considered missing if more than half of the items used to generate the domain score were unanswered. For single item symptoms, the score was considered missing if the item was unanswered. For each time point and each domain analysed, the proportion of scores 'missing' was computed by dividing the number of patients who did not have a domain/item score generated by the number of patients who completed a HRQL assessment at that time.

Both assessments of missing data were combined to provide a total % missing for each domain/item score generated, which represented the proportion of domain/items scores not completed or received out of those that were expected. This allowed a comparison of the total missing data between instrument groups.

3.5.2 Objective 2: Comparison between treatment arms by instrument

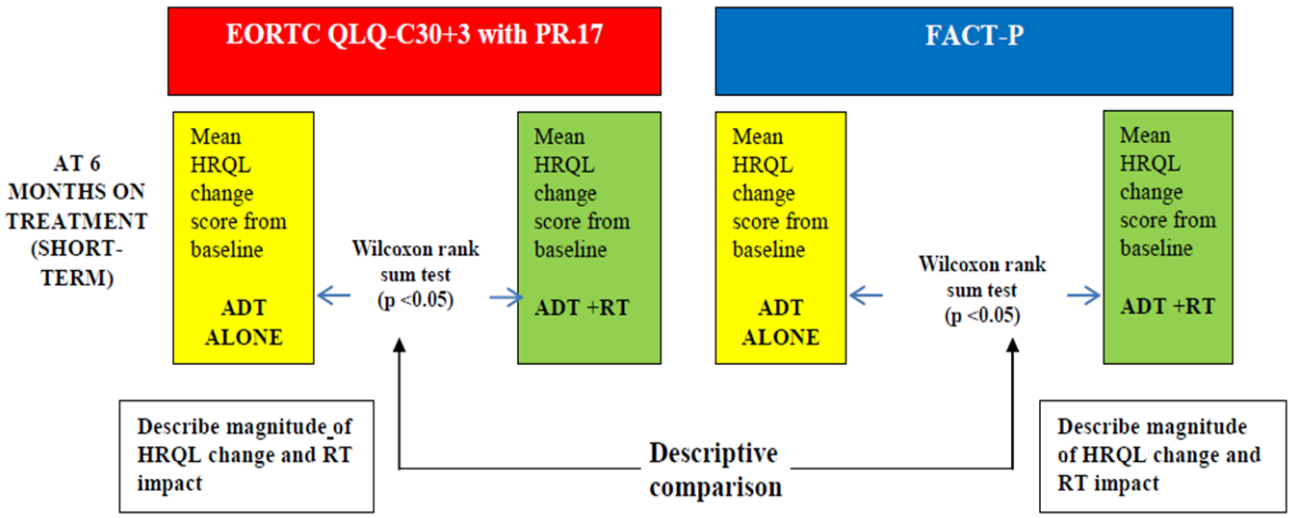
Change scores were generated for each participant at each assessment after baseline up to and including 36 months. Change scores represent the change from baseline, and thus are calculated

by subtracting each participant's baseline score from their score at each follow-up assessment. Each participant has a change score calculated for each assessment and for each domain/item. For symptom domains/items a negative change score indicates improvement from baseline (contrary to functional domains, in which a negative score indicates worsening).

The domain/item scores analyzed for this objective were those expected to be impacted by RT. For the EORTC-QLQ C-30+3 with PR17 this included urine symptoms (PR17), bowel/rectum symptoms (PR17), constipation (EORTC QLQ C30+3), diarrhea (EORTC QLQ C30+3), and urination at night (PR17). For the FACT-P, urine problems and bowel trouble were included. The 6 month assessment data was used for this objective. The mean and standard deviation for the change scores at 6 months was computed for each domain/item by treatment allocation and by instrument.

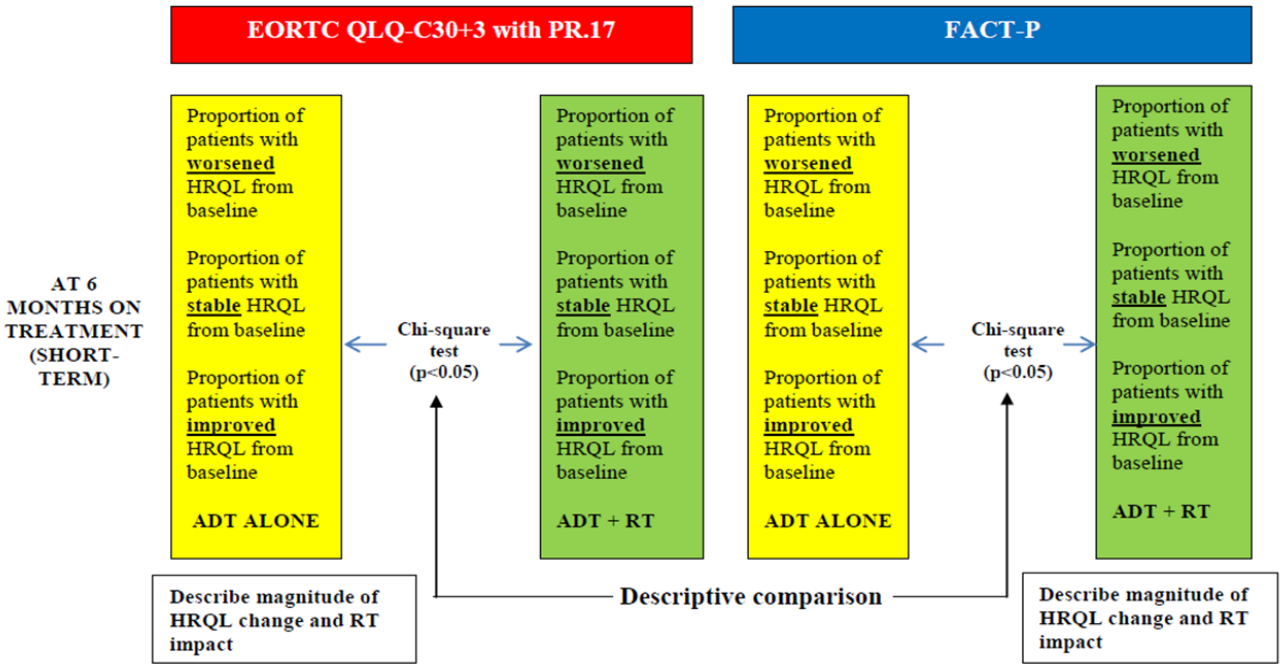
Figure 3.2 below depicts the comparison done for Objective 2A. A Wilcoxon rank sum test was performed to compare the two treatment arm's mean change scores. This was done separately for each instrument and for each 'comparable' domain/item score such that a descriptive comparison of the 2 instrument groups could be done.

Figure 3.2: Objective 2A analysis



For Objective 2B, the 6 month change scores were used to categorize each participant into 'worsened', 'stable' and 'improved' categories (indicating whether each patient was worse, stable or improved at 6 months as compared to baseline). This was done for each domain/item, using the criteria for clinically meaningful change as defined in section 3.4 (10 point change). Proportions in each category were compared by treatment arm using the chi square test. Again, this was done separately for each instrument and for each comparable domain/item score, such that a descriptive comparison of the 2 instrument groups could be performed, as indicated in Figure 3.3 below.

Figure 3.3: Objective 2B analysis

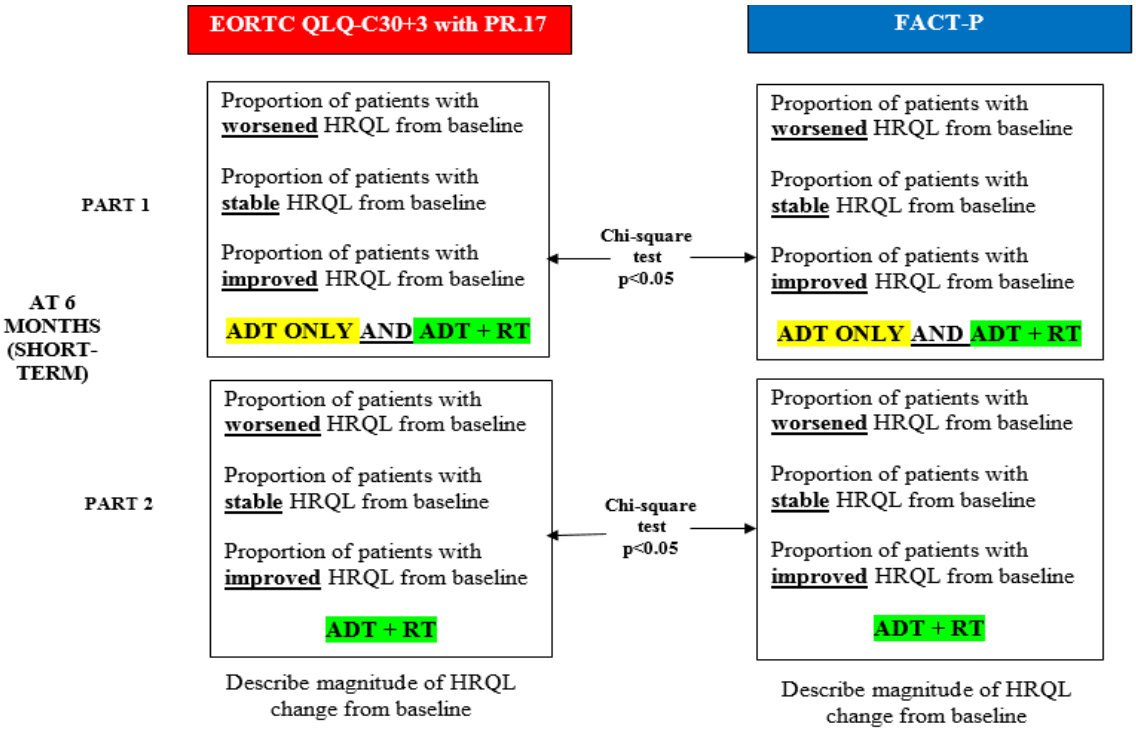


3.5.3 Objective 3: Comparison between Instruments

The purpose of Objective 3 was to directly compare the 2 instrument groups statistically, as opposed to Objective 2 which was a statistical comparison between treatment groups.

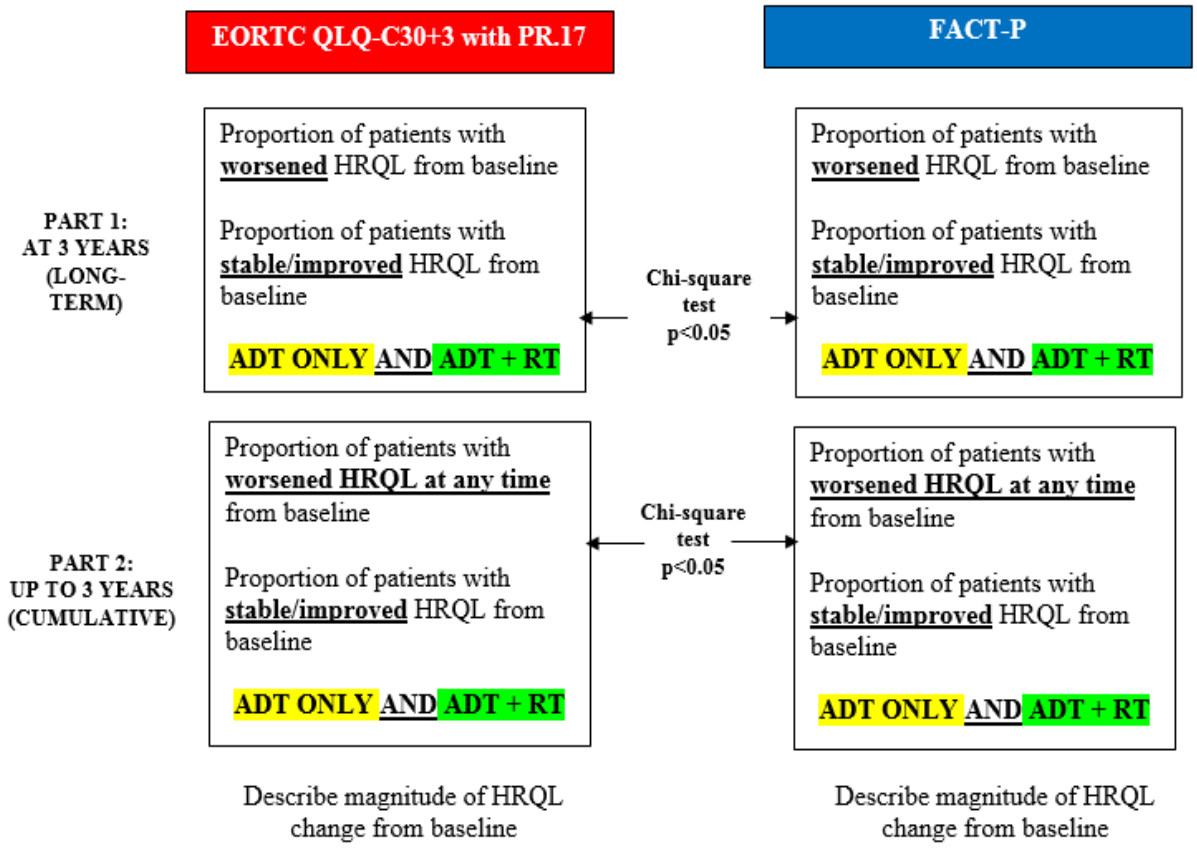
The 6 month assessment data was used for objective 3A. Based on the change scores from baseline, each participant was categorized into 'worsened', 'stable' and 'improved' categories (indicating if the patient was worse, stable or improved compared to baseline) for each domain/item, using the clinically meaningful change as defined in section 3.4 (10 point change). The domain/item scores analyzed were those expected to be impacted by RT in the short term, consistent with Objective 2. For the EORTC-QLQ C-30+3 with PR17 this included urine symptoms (PR17), bowel/rectum symptoms (PR17), constipation (EORTC QLQ C30+3), diarrhea (EORTC QLQ C30+3), and urination at night (PR17). For the FACT-P this included urine problems and bowel trouble. Figure 3.4 depicts the comparisons done in Objective 3A. Proportions in each category were compared by instrument group using the chi square test. The two instruments were compared with the treatment arms pooled (ADT and ADT +RT) and in the ADT + RT group only.

Figure 3.4: Objective 3A analysis



The analysis for Objective 3B is depicted in Figure 3.5. The domain/item scores analyzed for this objective were those expected to be impacted by ADT in the long term. For the EORTC-QLQ C-30+3 this included physical functioning, role functioning, emotional functioning, social functioning and fatigue. For the FACT-P this included physical well-being, functional well-being, emotional well-being, social/family well-being and fatigue. The analysis compared the 2 instrument groups, with treatment arm groups pooled together. Part 1 of the analysis used the 36 month assessment data. Change scores from baseline at 3 years were used to group patients into worsened or improved/stable from baseline. A patient was considered 'worsened' if their 36 month score was at least 10 points lower than their baseline score. Part 2 of the analysis utilized all HRQL data up until 3 years. Patients were grouped into the 'worsened' category if they had experienced a 10 point decline at any point in time from baseline to 36 months. Patients were considered stable/improved if they had not experienced a 10 point decline at any point up to 3 years. For Part 2, a sensitivity analysis was also performed in which the cut-point for clinically meaningful decline was changed from 10 points to 7 points.

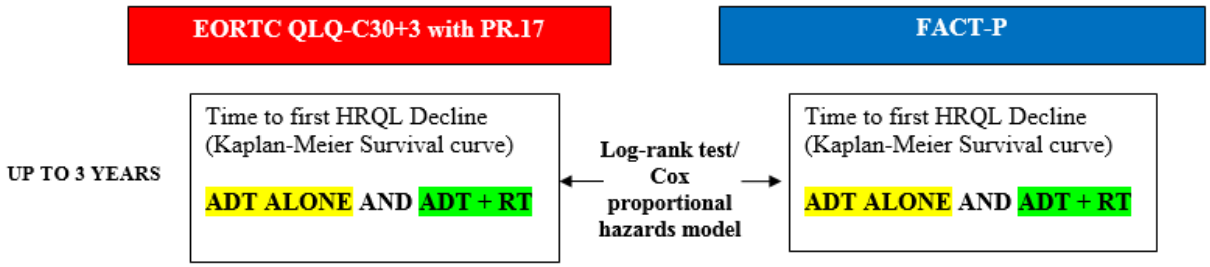
Figure 3.5: Objective 3B analysis



For Objective 3C, the time to clinically meaningful HRQL decline in the EORTC QLQ-C30+3/ PR17 and the FACT-P instrument groups were described using Kaplan-Meier survival curves, where a clinically meaningful decline was defined as a decrease of at least 10 points consistent with the other objectives. A log-rank test was performed to determine whether the two curves were statistically different. As randomization of instrument was performed by study site (not by patient), there was potential for some unbalance in baseline characteristics (including treatment allocation) between instrument groups, some of which could be predictors of clinically meaningful decline. Therefore a cox proportional hazards model was be applied to adjust for explanatory variables. Potential predictors of HRQL decline were first evaluated with a univariate

cox proportional hazards model. Those variables that were associated with HRQL decline at a p value of ≤0.2 were included in the final cox proportional hazards model. Potential covariates evaluated included age, treatment allocation, clinical stage, performance status, PSA level, type of hormonal therapy, method of lymph node staging, normal rectal exam, Gleason score and prior hormonal therapy. A sensitivity analysis for the time to clinically meaningful decline was also performed, adjusting the cut-off point for clinically meaningful decline from 10 points to 7 points. The intention of the sensitivity analysis was to test the robustness of the results obtained in the first part of the analysis, and to examine the extent to which the results are impacted by changing the assumption of what a clinically meaningful decline is. In this analysis, we assume a smaller change is considered clinically meaningful, thus potentially increasing the number of patients that meet the definition for clinically meaningful decline at any time point.

Figure 3.6: Objective 3C Analysis



Chapter 4

Results

4.1 Participant characteristics

The patient characteristics are displayed in Table 4.1 below, by instrument group. The median age of patients in this sample was 69.5 with the majority of patients over 60 years of age (94%). Most patients had an ECOG performance status of 0 at baseline (85%) and had clinical stage 3 prostate cancer (89%). Most patients had not had any prior hormonal therapy before study entry (89%). Very few patients chose an orchiectomy as the hormonal therapy on study, while most opted for the LHRH agonist (89%). Treatment arm allocation was relatively balanced between instrument groups, although the FACT-P group had slightly more patients in the ADT +RT arm (52%) compared to the EORTC QLQ C-30+3/PR17 group (48%). Other patient characteristics were for the most part similar in the two instrument groups. More patients had prior hormonal therapy in the FACT-P group (18%) vs. the EORTC QLQ-C30+3/PR17 group (6%). Also, more patients in the FACT-P group had clinical staging of lymph nodes (75%) compared to the EORTC QLQ-C30+3/PR17 group (66%), rather than radiological or surgical.

Table 4.1 Patient characteristics by instrument allocation

	EORTC QLQ-C30+3/PR17 N (%)		FACT-P N (%)		TOTAL N (%)	
TREATMENT ALLOCATION						
A (ADT)	93	(52)	64	(48)	157	(50)
B (ADT + RT)	86	(48)	68	(52)	154	(50)
METHOD OF LYMPH NODE STAGING						
Clinical	119	(66)	99	(75)	218	(70)
Radiological	57	(32)	25	(19)	82	(26)
Surgical	3	(2)	8	(6)	11	(4)

CLINICAL STAGE						
2	3	(2)	3	(2)	6	(2)
3a	74	(41)	72	(55)	146	(47)
3b	34	(19)	24	(18)	58	(19)
3c	46	(26)	24	(18)	70	(23)
4a	3	(2)	2	(2)	5	(2)
4b	7	(4)	5	(4)	12	(4)
NO	12	(7)	2	(2)	14	(5)
CHOICE OF HORMONAL THERAPY						
LHRH agonist	162	(90)	115	(87)	277	(89)
Bilateral Orchiectomy	17	(10)	17	(13)	34	(11)
PERFORMANCE STATUS (ECOG)						
0	150	(84)	115	(87)	265	(85)
1	29	(16)	16	(12)	45	(14)
2	0	(0)	1	(1)	1	(0)
PRIOR HORMONAL THERAPY						
N	169	(94)	108	(82)	277	(89)
Y	10	(6)	24	(18)	34	(11)
RECTAL EXAM NORMAL						
N	160	(89)	115	(87)	275	(88)
Y	19	(11)	17	(13)	36	(12)
AGE AT ALLOCATION						
Median	69.6		69.5		69.5	
40-49	1	(1)	0	(0)	1	(0)
50-59	10	(6)	7	(5)	17	(5)
60-69	83	(46)	63	(48)	146	(47)
>=70	85	(47)	62	(47)	147	(47)
PSA						
Median	25.1		24.7		25.0	
<20	77	(43)	53	(40)	130	(42)
20-50	59	(33)	52	(39)	111	(36)
>50	43	(24)	27	(20)	70	(23)

GLEASON SCORE							
Median	7		7		7		
<8	114	(64)	77	(58)	191	(61)	
8-10	65	(36)	55	(42)	120	(39)	
TOTAL	179	(100)	132	(100)	311	(100)	

4.2 Objective 1

4.2.1 Mean HRQL scores by instrument and treatment arm

Table 4.2 below shows the mean HRQL scores at baseline, 6 months and 36 months for the selected domains/items, by treatment allocation for the EORTC QLQ-C30 + PR17 questionnaire. In terms of the functional domains, physical functioning, role functioning and social functioning mean scores decreased over time in both treatment groups (as indicated by lower mean scores), whereas emotional functioning appeared to remain relatively stable. Fatigue also worsened over time in both treatment groups (as indicated by higher mean scores), consistent with known ADT effects. Bowel/rectum symptoms and diarrhea worsened over time in the ADT + RT group only, consistent with radiation treatment effects, with mean scores highest at 6 months. Urine symptoms appeared to improve slightly over time in the ADT group but remained relatively stable in the ADT +RT group, according to the mean scores. Urination at night, however, worsened over time in both treatment groups.

Table 4.2 EORTC QLQ-C30+3/PR.17 Mean HRQL Scores

		Baseline		6 months		36 months	
		ADT	ADT + RT	ADT	ADT + RT	ADT	ADT + RT
Physical Functioning	N	93	86	76	75	66	61
	Missing (N, (%))	0	0	1 (1.3)	0	1 (1.5)	0
	Mean	92.47	91.59	89.29	87.17	83.69	82.93
	Std dev	12.02	16.00	17.32	19.79	20.03	21.28
Role Functioning	N	92	86	76	75	65	60
	Missing (N, (%))	1 (1.1)	0	1 (1.3)	0	2 (3.0)	1 (1.6)
	Mean	94.93	94.77	91.67	87.56	84.10	83.06

	Std dev	11.78	13.83	15.99	23.90	21.33	27.19
Emotional Functioning	N	91	86	76	75	66	61
	Missing (N, (%))	2 (2.1)	0	1 (1.3)	0	1 (1.5)	0
	Mean	85.16	83.04	86.07	87.11	84.39	84.65
	Std dev	14.29	17.82	16.24	14.45	15.69	15.95
Social Functioning	N	91	86	76	73	64	59
	Missing (N, (%))	2 (2.1)	0	1 (1.3)	2 (2.7)	3 (4.5)	2 (3.3)
	Mean	95.05	94.19	89.91	87.67	82.81	82.20
	Std dev	12.30	14.41	21.44	21.16	26.05	28.85
Urine Symptoms	N	92	86	74	75	66	61
	Missing (N, (%))	1 (1.1)	0	3 (3.9)	0	1 (1.5)	0
	Mean	9.55	11.63	7.89	11.49	7.31	10.90
	Std dev	9.38	13.01	9.96	9.85	8.37	14.63
Bowel/ Rectum Symptoms	N	90	85	76	75	65	61
	Missing (N, (%))	3 (3.2)	1 (1.2)	1 (1.3)	0	2 (3.0)	0
	Mean	3.70	3.33	1.54	6.44	3.08	5.74
	Std dev	7.40	8.45	6.19	13.94	9.27	12.50
Fatigue	N	93	86	76	75	66	61
	Missing (N, (%))	0	0	1 (1.3)	0	1 (1.5)	0
	Mean	14.04	14.08	17.91	21.85	22.73	23.68
	Std dev	13.99	17.99	21.23	21.58	20.20	23.08
Constipation	N	91	86	76	75	65	61
	Missing (N, (%))	2 (2.1)	0	1 (1.3)	0	2 (3.0)	0
	Mean	6.96	6.59	11.40	6.67	4.62	6.67
	Std dev	18.93	15.18	24.07	16.44	14.28	14.78
Diarrhea	N	92	85	76	75	65	60
	Missing (N, (%))	1 (1.1)	1 (1.2)	1 (1.3)	0	2 (3.0)	1 (1.6)
	Mean	4.35	5.88	2.63	13.33	4.41	6.67
	Std dev	11.29	17.20	9.05	20.50	13.99	14.67
Urination at Night	N	90	84	75	74	58	60
	Missing (N, (%))	3 (3.2)	2 (2.3)	2 (2.6)	1 (1.3)	9 (13.4)	1 (1.6)
	Mean	37.41	37.70	42.67	47.30	44.83	47.78
	Std dev	25.39	23.01	26.03	20.65	25.40	24.83

Table 4.3 shows the mean HRQL scores at baseline, 6 months and 36 months for the selected domains/items, by treatment allocation for the FACT-P questionnaire. Consistent with

the EORTC QLQ-C30+3, the FACT-P showed decline over time in both treatment groups for all functional domains except for emotional functioning which appeared to improve over time based on mean scores. Fatigue also worsened over time in both treatment groups, as expected. Urine problems, according to mean scores, appeared to improve over time in both treatment groups. Bowel trouble showed worsening in the ADT group and improvement in the ADT + RT group**.**

Table 4.3 FACT-P Mean HRQL Scores

		Baseline		6 months		36 months	
		ADT	ADT + RT	ADT	ADT + RT	ADT	ADT + RT
Physical Well-being	N	64	68	58	53	38	37
	Missing (N, (%))	0	0	1 (1.7)	3 (5.3)	3 (7.3)	4 (9.7)
	Mean	91.19	89.36	85.92	83.75	82.98	85.47
	Std dev	9.86	15.17	13.57	15.76	15.19	13.67
Functional Well-being	N	64	68	57	52	38	37
	Missing (N, (%))	0	0	2 (3.4)	4 (7.1)	3 (7.3)	4 (9.7)
	Mean	82.98	78.76	78.32	76.69	73.97	75.95
	Std dev	16.02	19.90	16.89	17.54	17.65	18.99
Emotional Well-being	N	64	68	57	52	38	37
	Missing (N, (%))	0	0	2 (3.4)	4 (7.1)	3 (7.3)	4 (9.7)
	Mean	81.88	81.53	87.48	87.98	86.91	88.07
	Std dev	16.80	16.39	13.35	12.42	12.67	16.50
Social Well-being	N	64	68	58	53	38	37
	Missing (N, (%))	0	0	1 (1.7)	3 (5.3)	3 (7.3)	4 (9.7)
	Mean	81.56	82.55	80.06	81.35	74.19	79.13
	Std dev	14.98	15.77	15.12	16.96	18.16	17.22
Urine Problems	N	64	68	58	52	38	37
	Missing (N, (%))	0	0	1 (1.7)	4 (7.1)	3 (7.3)	4 (9.7)
	Mean	32.75	25.37	24.71	22.76	26.10	23.09
	Std dev	27.66	23.76	22.78	22.27	23.34	17.69
Bowel Trouble	N	64	68	57	52	38	36
	Missing (N, (%))	0	0	2 (3.4)	4 (7.1)	3 (7.3)	5 (13.9)
	Mean	11.72	14.34	13.16	11.54	15.79	11.11
	Std dev	21.81	27.08	23.67	25.94	28.13	21.08
Fatigue	N	64	68	58	53	38	37
	Missing (N, (%))	0	0	1 (1.7)	3 (5.3)	3 (7.3)	4 (9.7)
	Mean	24.61	29.04	37.07	36.32	40.79	39.19
	Std dev	27.27	27.86	30.79	33.46	28.72	29.19

4.2.2 Missing data

In some cases patients submitted a questionnaire but did not answer certain items within the questionnaires such that a domain score could not be generated (i.e. was considered 'missing'). If more than half of the items that make up a domain were not completed, the score was considered missing. In tables 4.2 and 4.3 above, for each domain/item score, the N refers to the number of non-missing values (i.e. the number of patients that had sufficient number of responses to generate the domain/item score). The numbers of 'Missing' scores are also displayed, representing the number of patients that did not provide enough answered items to calculate a given domain/item scores.

Missing data was also assessed by evaluating the questionnaire compliance. Table 4.4 shows compliance at the scheduled assessment time points from baseline to 3 years for both instruments. The number of expected questionnaires represents the number of participants who were continuing on the study at each scheduled visit. Compliance was quite good up to 3 years, and was similar between the 2 instrument groups (ranged from 86.7% to 94.9% in the EORTC group and 88.6% to 95.5% in the FACT group).

Table 4.4: Compliance with HRQL Questionnaire Completion by Instrument Allocation

	baseline	6 mos	12 mos	18 mos	24 mos	36 mos
EORTC QLQ-C30 +PR.17						
Expected (N)	179	166	158	151	143	136
Received (N, %)	179 (100)	152 (91.6)	150 (94.9)	137 (90.7)	124 (86.7)	128 (94.9)
FACT-P						
Expected (N)	132	120	116	105	98	92
Received (N, %)	132 (100)	115 (95.8)	106 (91.4)	93 (88.6)	87 (88.8)	82 (92.9)

To get a sense of the total missing data, it was necessary to factor in the compliance (received/expected) as shown in table 4.4, as well as the missing domain scores attributed to missing items (from tables 4.2 and 4.3). Table 4.5 and 4.6 displays the total missing data for each domain/item score analyzed (i.e. the total missing domain scores including those that did not submit a questionnaire when it was expected, and those that submitted a questionnaire but did not complete enough items to generate the domain/item score) for each instrument. Missing data was similar between the 2 instruments at 6 months, ranging from 9% to 10.8% for EORTC QLQ-C30+3/PR.17 and 7.5% to 9.2% for the FACT-P. However, at 36 months there was more data missing overall in the FACT-P group (18.5% to 19.6%) compared to the EORTC QLQ-C30+3/PR.17 group (6.6% to 13.2%).

Table 4.5: EORTC QLQ-C30+3/PR.17 Total Missing Data

Domain/Item		Baseline	6 months	36 months
Physical Functioning	N (missing questionnaires)	0	14	8
	N (missing scores)	0	4	1
	N (expected)	179	166	136
	Total % missing	0	10.8	6.6
Role Functioning	N (missing questionnaires)	0	14	8
	N (missing scores)	1	1	3
	N (expected)	179	166	136
	Total % missing	0.6	9	8.1
Emotional Functioning	N (missing questionnaires)	0	14	8
	N (missing scores)	2	1	1
	N (expected)	179	166	136
	Total % missing	1.1	9	6.6
Social Functioning	N (missing questionnaires)	0	14	8
	N (missing scores)	2	3	5
	N (expected)	179	166	136
	Total % missing	1.1	10.2	9.6
Urine Symptoms	N (missing questionnaires)	0	14	8
	N (missing scores)	1	3	1
	N (expected)	179	166	136
	Total % missing	0.6	10.2	6.6

Bowel/ Rectum Symptoms	N (missing questionnaires)	0	14	8
	N (missing scores)	4	1	2
	N (expected)	179	166	136
	Total % missing	2.2	9	7.3
Fatigue	N (missing questionnaires)	0	14	8
	N (missing scores)	0	1	1
	N (expected)	179	166	136
	Total % missing	0	9	6.6
Constipation	N (missing questionnaires)	0	14	8
	N (missing scores)	2	1	2
	N (expected)	179	166	136
	Total % missing	1.1	9	7.3
Diarrhea	N (missing questionnaires)	0	14	8
	N (missing scores)	2	1	3
	N (expected)	179	166	136
	Total % missing	1.1	9	8.1
Urination at Night	N (missing questionnaires)	0	14	8
	N (missing scores)	5	3	10
	N (expected)	179	166	136
	Total % missing	2.8	10.2	13.2

Table 4.6: FACT-P Total Missing Data

Domain/Item		Baseline	6 months	36 months
Physical Well-being	N (missing questionnaires)	0	5	10
	N (missing scores)	0	4	7
	N (expected)	132	120	92
	Total % missing	0	7.5	18.5
Functional Well-being	N (missing questionnaires)	0	5	10
	N (missing scores)	0	6	7
	N (expected)	132	120	92
	Total % missing	0	9.2	18.5
Emotional Well-being	N (missing questionnaires)	0	5	10
	N (missing scores)	0	6	7
	N (expected)	132	120	92
	Total % missing	0	9.2	18.5
Social Well-being	N (missing questionnaires)	0	5	10
	N (missing scores)	0	4	7
	N (expected)	132	120	92
	Total % missing	0	7.5	18.5

Urine Problems	N (missing questionnaires)	0	5	10
	N (missing scores)	0	5	7
	N (expected)	132	120	92
	Total % missing	0	8.3	18.5
Bowel Trouble	N (missing questionnaires)	0	5	10
	N (missing scores)	0	6	8
	N (expected)	132	120	92
	Total % missing	0	9.2	19.6
Fatigue	N (missing questionnaires)	0	5	10
	N (missing scores)	0	4	7
	N (expected)	132	120	92
	Total % missing	0	7.5	18.5

4.3 Objective 2

4.3.1 Comparison of mean change scores between treatment arms

Table 4.7 below shows the mean change scores from baseline at 6 months and the comparison between treatment arms, for each instrument and for the selected symptom domains/items.

Urine symptoms:

The PR17 checklist (administered with the EORTC QLQ-C30+3) includes a urine symptom domain, as does the FACT-P prostate cancer subscale. The mean change scores for the PR17 showed overall slight improvement in urine symptoms from baseline in both the ADT only and ADT +RT treatment groups (-1.33 and -0.18 respectively). Although there was slightly more improvement in the ADT arm compared to the ADT + RT arm, this difference was not statistically significant (p=0.107). The FACT-P, however, showed overall slight worsening in the ADT+RT arm based on mean change score (1.12), versus the ADT only arm which showed overall improvement (-9.27). For the FACT-P, the difference between treatment arms for urine symptom mean change scores was statistically significant (p=0.0185).

Bowel/rectum symptoms/trouble:

Mean change scores indicated bowel/rectum symptoms and diarrhea were improved in the ADT arm (-1.35 and -1.78 respectively) and worsened in the ADT+RT arm (3.33 and 7.66 respectively) according to the EORTC QLQ-C30+3/PR17. The difference between treatment arms for both of these symptoms were statistically significant, according the EORTC QLQ-C30+3/PR17 (p=.0191 and p= 0.0005, respectively). The FACT-P, while it does include 1 question regarding bowel trouble, does not include an item or domain for diarrhea and does not adequately capture the rectum/bowel symptoms relevant to RT effects. Contrary to the EORTC QLQ-C30+3/PR17, FACT-P bowel trouble mean change score showed worsening in the ADT only arm (1.75) and improvement in the ADT+RT (-2.40). The FACT-P did not find a significant between-treatment arm difference in mean change scores for this 1 bowel item.

Table 4.7 – HRQL Changes from baseline at 6 months by treatment arm for symptom domains/items

	ADT			ADT + RT			
	N	Mean	SD	N	Mean	SD	P
EORTC QLQ-C30+3/PR.17							
urine symptoms	74	-1.33	10.90	75	-0.18	13.69	0.107
Bowel/rectum symptoms	74	-1.35	7.17	75	3.33	14.76	0.0191
constipation	74	4.50	18.57	75	0.00	16.44	0.1854
diarrhea	75	-1.78	12.12	74	7.66	21.76	0.0005
urination at night	73	7.31	25.61	74	9.91	28.53	0.4617
FACT-P							
urine symptoms	58	-9.27	20.16	52	1.12	23.86	0.0185
bowel trouble	57	1.75	16.27	52	-2.40	25.37	0.3062

4.3.2 Comparison of % improved, stable and worsened between treatment arms

Table 4.8 below shows the proportions of patients with improved, stable and worsened scores at 6 months. Treatment arms are again compared in each instrument group, for the same symptom items/domains as in table 4.7.

Urine symptoms:

The majority of patients remained stable from baseline, in both the ADT and ADT+RT groups. However, more patients had worsened urine scores in the ADT+RT (16%) group compared to the ADT group (5%) according to the EORTC QLQ-C30+3/PR17. The between-treatment arm difference in proportions improved, stable and worsened, however was not statistically significant for the EORTC QLQ-C30/PR17 (p=0.1129). Similarly, according to the FACT-P, the ADT+RT group had a greater proportion worsened (27%) compared to the ADT only group (16%) but the difference was not statistically significant (p=0.0709).

Bowel symptoms:

A considerably greater proportion of patients were worsened in the ADT+RT group compared to the ADT only group according the EORTC QLQ-C30+3/PR17 for bowel/rectum symptoms (20% and 5%, respectively) and for diarrhea (32% and 4%, respectively). Consistent with the analysis of mean change scores (Table 4.7), the EORTC QLQ-C30+3/PR17 detected statistically significant between-treatment arm differences in bowel/rectum symptoms (p=0.0257) and diarrhea (p<.0001). The FACT-P did not detect significant between-treatment arm differences in the one bowel trouble item.

Table 4.8 HRQL Response at 6 months by treatment arm for symptom domains/items

	ADT				ADT + RT				
	Total N	Improved N (%)	Stable N (%)	Worsened N (%)	Total N	Improved N (%)	Stable N (%)	Worsened N (%)	P
EORTC QLQ-C30+3/PR17									
Urine symptoms	74	11 (15)	59 (80)	4 (5)	75	10 (13)	53 (71)	12 (16)	0.1129
Bowel/rectum symptoms	74	10 (13)	60 (81)	4 (5)	75	7 (9)	53 (71)	15 (20)	0.0257
Constipation	74	5 (7)	56 (76)	13 (18)	75	7 (9)	60 (80)	8 (11)	0.4371
Diarrhea	75	7 (9)	65 (87)	3 (4)	74	6 (8)	45 (61)	23 (32)	<.0001
Urination at night	73	9 (12)	44 (60)	20 (27)	74	11 (15)	35 (47)	28 (38)	0.2792
FACT-P									
Urine symptoms	58	22 (38)	27 (47)	9 (16)	52	10 (19)	28 (54)	14 (27)	0.0709
Bowel trouble	57	6 (11)	42 (74)	9 (16)	52	10 (19)	35 (67)	7 (13)	0.4359

4.4 Objective 3

Objective 3 focused on directly comparing the 2 HRQL instrument groups with statistical testing.

4.4.1 Between-instrument comparison of % improved, stable and worsened (symptoms)

Table 4.9 and 4.10 below display the proportions of patients improved, stable and worsened at 6 months in each instrument group for the symptom domains/items and the comparison between the 2 instrument groups. Each symptom domain/item from the EORTC QLQ-C30+30/PR17 was compared to a symptom domain/item on the FACT-P. Table 4.9 includes all patients (treatment groups pooled) and Table 4.10 shows the same comparison but with the sample restricted to the ADT + RT arm.

Urine symptoms:

When the treatment groups are combined (table 4.9), there are significant differences observed between instruments in terms of the proportions improved, stable and worsened for urine symptoms at 6 months (p=0.0002). The FACT-P reported a greater proportion of patients with improved scores (29%) and a greater proportion of patients with worsened scores (21%) compared to the PR17 (14% and 11%, respectively). The PR17 reported a greater proportion of patients with stable urine symptom scores (75%) compared to the FACT-P (50%). The PR17 also contains a 'urination at night' single item, which when compared with the FACT-P urine symptom domain was also significantly different in terms of proportions of patients improved, stable and worsened (p=0.0046). Again, the FACT-P showed more worsening and more improvement compared to the PR17 which showed more patients with stable symptom scores.

However, between-instrument group differences in proportions of patients with improved, stable and worsened urine symptom scores at 6 months were no longer statistically significant when the sample was restricted to the ADT +RT arm, as shown in table 4.10 (p=0.1473). Consistent with table 4.9, however, the FACT-P shows a greater proportion of patients with improved or worsened symptoms compared to the EORTC QLQ-C30+3/PR17.

Restricting the sample to include only the ADT+RT arm resulted in both instruments showing a greater proportion of patients with worsened urine symptoms compared to when the treatment groups were combined.

Bowel/rectum symptoms/trouble:

The EORTC QLQ-C30+3/PR17 constipation, diarrhea and bowel/rectum symptoms were compared against the one bowel trouble item on the FACT-P. When the treatment groups were combined, no statistically significant differences were observed between the 2 instruments for any of the comparisons of proportions improved, stable and worsened for these domains. There were also no notable trends by instrument.

However when the sample was restricted to the ADT+RT arm, a statistically significant difference was observed between the EORCT QLQ C-30+3 diarrhea and the FACT-P bowel trouble (p=0.0279), with EORTC QLQ C-30+3 reporting a greater proportion of patients with worsened diarrhea symptom scores (31%) compared to the FACT-P bowel trouble (13%).

Table 4.9 HRQL response at 6 months by instrument allocation for symptom domains/items (treatment arms pooled)

	EORTC QLQ-C30+3				FACT-P				
	Total	Improved	Stable	Worsened	Total	Improved	Stable	Worsened	
	N	N (%)	N (%)	N (%)	N	N (%)	N (%)	N (%)	P
EORTC constipation/ FACT bowel trouble	149	12 (8)	116 (78)	21 (14)	109	16 (15)	77 (71)	16 (15)	0.2233
EORTC diarrhea /FACT bowel trouble	149	13 (8)	110 (74)	26 (17)	109	16 (15)	77 (71)	16 (15)	0.3057
EORTC rectum/bowel symptoms/ FACT bowel trouble	149	17 (11)	113 (76)	19 (13)	109	16 (15)	77 (71)	16 (15)	0.6284

EORTC urine symptoms/ FACT urine symptoms	149	21 (14)	112 (75)	16 (11)	110	32 (29)	55 (50)	23 (21)	0.0002
EORTC urination at night/ FACT urine symptoms	147	20 (14)	79 (54)	48 (33)	110	32 (29)	55 (50)	23 (21)	0.0046

Table 4.10 HRQL response at 6 months by instrument allocation for symptom domains/items (ADT+RT Arm only)

	EORTC QLQ-C30+3				FACT-P				
	Total	Improved	Stable	Worsened	Total	Improved	Stable	Worsened	
	N	N (%)	N (%)	N (%)	N	N (%)	N (%)	N (%)	P
EORTC constipation/ FACT bowel trouble	75	7 (9)	60 (80)	8 (11)	52	10 (19)	35 (67)	7 (13)	0.211
EORTC diarrhea /FACT bowel trouble	74	6 (8)	45 (61)	23 (31)	52	10 (19)	35 (67)	7 (13)	0.0279
EORTC rectum/bowel symptoms/ FACT bowel trouble	75	7 (9)	53 (71)	15 (20)	52	10 (19)	35 (67)	7 (13)	0.2171
EORTC urine symptoms/ FACT urine symptoms	75	10 (13)	53 (71)	12 (16)	52	10 (19)	28 (54)	14 (27)	0.1473
EORTC urination at night/ FACT urine symptoms	74	11 (15)	35 (47)	28 (38)	52	10 (19)	28 (54)	14 (27)	0.4268

4.4.2 Between-instrument comparison of % improved/stable and worsened (functional domains and fatigue)

Table 4.11 shows the proportions of patients with improved or stable and worsened HRQL from baseline *at* the 36 month assessment in each instrument group for the four common functional domains and fatigue, and the comparison between the 2 instrument groups. Table 4.12 shows the proportions of patients improved or stable and worsened at any point in time *up to and including* 36 months. In the latter analysis, patients are categorized into the 'worsened' group if at

any point they experienced worsening (i.e. at least 10 point decline) from baseline up to and including the 36 month assessment. Patients in the 'stable/improved' category are those that did not experience worsening at any point (i.e. remained stable or improved for the entire follow up period up until 3 years).

Physical functioning/well-being:

At the 36 month HRQL assessment, the majority of patients remained stable or improved from baseline in terms of physical functioning/well-being. However, the EORTC QLQ-C30+3 did report a greater proportion of patients with worsened physical functioning at this time point compared to the FACT-P (43% vs 33%). In the cumulative analysis, however, the FACT-P reported that slightly more patients had worsened at some point in time from baseline to 36 months compared to the EORTC QLQ-C30+3 (59% vs 52%). However, neither of these between-instrument differences were statistically significant (p=0.1961 and p=0.2264 respectively).

Social functioning/well-being

The FACT-P showed a slightly greater proportion of patients with worsened social functioning scores (40%) compared to the EORTC QLQ-C30+3 (36%) at the 36 month assessment. This was also the case in the cumulative analysis (56% vs 55%) but the between-instrument group differences were not statistically significant (p=0.6098 and p=0.8725).

Emotional functioning/well-being

Compared to the other functional domains, there were considerably less patients worsened in the emotional domain, both at the 36 month assessment and at any point up to and including 36 months. While the EORTC QLQ-C30+3 showed a slightly greater proportion of patients with worsened emotional functioning at 36 months compared to the FACT-P, the FACT-P reported a slightly greater proportion of patients with worsened emotional functioning at any

point in time up to and including 36 months. Between-instrument group differences were not statistically significant for either analysis.

Role functioning/functional well-being

At the 36 month HRQL assessment, the EORTC QLQ-C30+3 reported a slightly greater proportion of patients with worsened HRQL than the FACT-P in terms of role functioning/functional well-being (36% vs 33%, p=0.6721). However the FACT-P reported a greater proportion of patients with worsened functional well-being at any time compared to the EORTC QLQ-C30's role functioning (60% vs 52%, p=0.1823). Again, the between-instrument group differences in proportions were not statistically significant.

Fatigue

Compared to the functional domains, fatigue showed the most clinically meaningful worsening both at 36 months and at any point in time up to and including 36 months. At 36 months, 59% of patients had worsened fatigue according to the FACT-P and 54% had worsened fatigue according to the EORTC QLQ-C30+3. In the cumulative analysis, 71% had worsened fatigue at some point according to the FACT-P compared to 75% according to the EORTC QLQ-C30+3. No statistically significant between-instrument group differences in proportions were observed for fatigue.

Table 4.11 HRQL response at 36 months by instrument allocation for functional domains and fatigue

	EORTC			FACT			
	Total N	Improved/Stable N (%)	Worsened N (%)	Total N	Improved/Stable N (%)	Worsened N (%)	P
physical functioning/ physical well-being	127	73 (57)	54 (43)	75	50 (67)	25 (33)	0.1961
Social functioning/ Social Well being	121	77 (64)	44 (36)	75	45 (60)	30 (40)	0.6098

Emotional functioning/ Emotional well-being	125	103 (82)	22 (18)	75	63 (84)	12 (16)	0.7706
Role functioning/ Functional well-being	124	79 (64)	45 (36)	75	50 (67)	25 (33)	0.6721
Fatigue	127	58 (46)	69 (54)	75	31 (41)	44 (59)	0.5487

Table 4.12 Cumulative HRQL response up to and including 36 months by instrument allocation for functional domains and fatigue

	EORTC QLQ-C30+3			FACT-P			
	Total N	Improved/Stable N (%)	Worsened N (%)	Total N	Improved/Stable N (%)	Worsened N (%)	P
physical functioning/ physical well-being	173	83 (48)	90 (52)	127	52 (41)	75 (59)	0.2264
Social functioning/ Social Well being	171	77 (45)	94 (55)	127	56 (44)	71 (56)	0.8725
Emotional functioning/ Emotional well-being	171	124 (73)	47 (27)	126	87 (69)	39 (31)	0.515
Role functioning/ Functional well-being	172	83 (48)	89 (52)	126	51 (40)	75 (60)	0.1823
Fatigue	173	44 (25)	129 (75)	127	37 (29)	90 (71)	0.4757

In summary, while there were some slight between-instrument group differences observed in terms of proportions improved/stable and worsened for the functional domains at 36 months and cumulatively up to and including 36 months as reported by either the EORTC QLQ-C30+3 or FACT-P, none of these differences were statistically significant.

Sensitivity analysis

A sensitivity analysis was performed for the cumulative response analysis, altering the cut-off for clinically meaningful change from 10% (10 points) to 7% (7 points). Results of this analysis are displayed in table 4.13.

There was no change in proportions for the EORTC QLQ-C30+3's social and role domains as well as fatigue, when a 7% CMC cut-point was used compared with the 10% cut-point. For the physical domain, one 1 more patient became worsened when the CMC cut-point was reduced. There was a substantial increase in the number of patients worsened only for the emotional domain. For the FACT-P however, a greater proportion of patients became worsened when the CMC cut-point was reduced to 7% for the physical, social, and role functioning domains. The proportions of patients improved/stable and worsened did not change for the FACT-P emotional and fatigue domains.

The between-instrument group differences in proportions of patients improved/stable and worsened became statistically significant for the physical, emotional and role/functional domains, using the 7% cut-off criteria ($p=0.0023$, $p=0.0044$ and $p=0.001$, respectively). The overall trend remained consistent in that for the functional domains the FACT-P showed a greater proportion of patients worsened at any point up to and including 36 months compared to the EORTC QLQ-C30+3, with the exception of emotional functioning/well-being. For the emotional domain, the EORTC QLQ-C30+3 showed significantly more patients worsened compared to the FACT-P ($p=0.0044$).

Table 4.13 Cumulative HRQL response up to and including 36 months by instrument allocation for functional domains and fatigue (sensitivity analysis)

	EORTC QLQ-C30+3			FACT-P			
	Total N	Improved/Stable N (%)	Worsened N (%)	Total N	Improved/Stable N (%)	Worsened N (%)	P
physical functioning/ physical well-being	173	82 (47)	91 (53)	127	38 (30)	89 (70)	0.0023
Social functioning/ Social Well being	171	77 (45)	94 (55)	127	47 (37)	80 (63)	0.1648
Emotional functioning/ Emotional well-being	171	90 (53)	81 (47)	126	87 (69)	39 (31)	0.0044
Role functioning/ Functional well-being	172	83 (48)	89 (52)	126	37 (29)	89 (71)	0.001
Fatigue	173	44 (25)	129 (75)	127	37 (29)	90 (71)	0.4757

4.4.3 Comparison of time- to-clinically meaningful HRQL decline between instruments

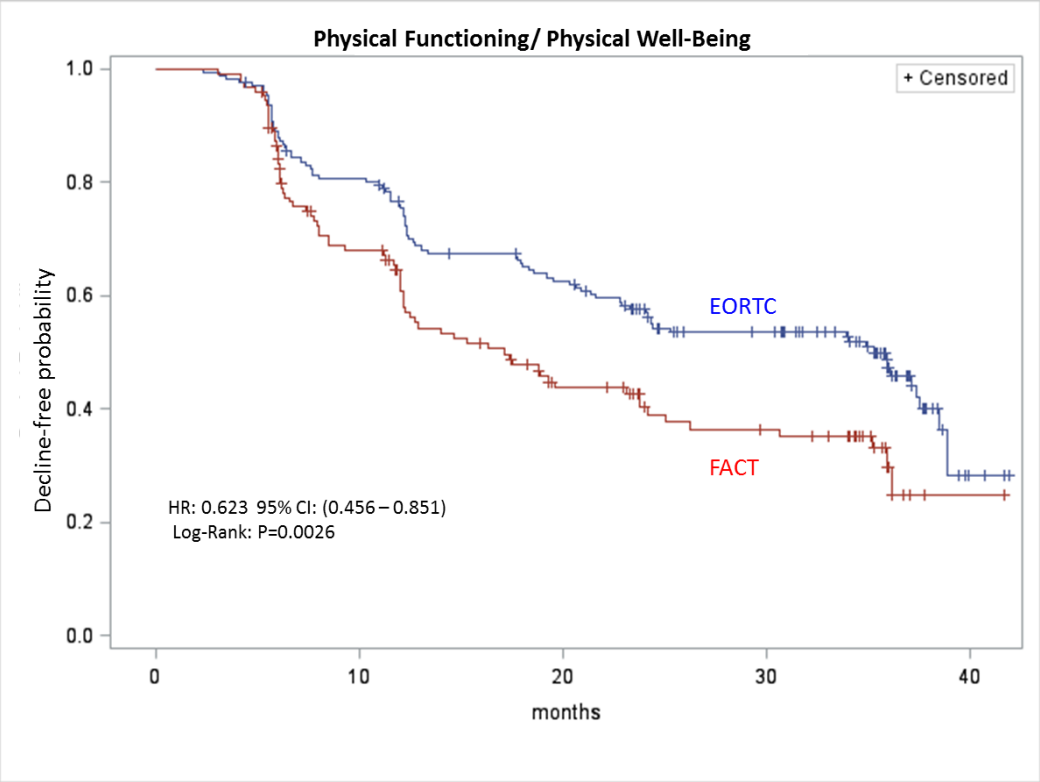
The Kaplan-Meier curves for time to HRQL decline for the comparable functional domains and fatigue, stratified by Instrument group, are displayed in figures 4.1 to 4.5. The FACT-P reported decline in physical functioning (log-rank p=0.0026) and role functioning (log-rank p=0.0137) earlier than the EORTC QLQ-C30+3. Rates of decline in the other functional domains and fatigue were comparable based on the reporting of the FACT-P or the EORTC QLQ C30+3.

Time-to-HRQL decline in physical functioning/physical well-being

For the physical functioning/physical well-being domains (Figure 4.1) there was a significant difference in the rate of decline between the EORTC QLQ-C30+3 and FACT-P instrument groups (log-rank p=0.0026). The separation of the 2 Kaplan Meier curves is apparent at approximately 6 months, the time of the first HRQL assessment, and this trend persisted over

time. The mean time-to-decline was 26.1 months for the EORTC QLQ-C30+3 group compared to 20.3 months for the FACT-P group, thus patients completing the FACT-P instrument showed higher rate of decline compared to those completing the EORTC QLQ-C30+3.

Figure 4.1: Time-to-HRQL decline for physical functioning/physical well-being by instrument allocation



Crude and adjusted associations between instrument and physical functioning/physical well-being

Table 4.14 displays the crude and adjusted hazard ratios and 95% confidence intervals for instrument allocation and other potential predictive factors for HRQL decline. The unadjusted hazard ratio (HR) for instrument was 0.623 indicating that, compared to the FACT-P group, the EORTC showed a decreased rate of decline in physical functioning/well-being HRQL. Out of the ten potential covariates evaluated PSA, age and Gleason score was associated with decline in physical functioning/well-being (at a p value of ≤ 0.2). Participants who had a PSA of 20-50 or

<20 at baseline had a decreased rate of decline in physical functioning/well-being then those that had a PSA>50, and this was statistically significant in the <20 group (HR=0.566, p=0.0038). Participants aged 60-69 had an increased rate of decline and participants 50-59 had a reduced rate of decline compared to the >=70 group, however the hazard ratios were not significant in the unadjusted or adjusted models. Participants with a Gleason score of 8-10 also had reduced rate of decline compared to participants with Gleason score <8, but this was not significant. It is notable that there was a slight imbalance between instrument groups for baseline PSA levels and Gleason score, however age was quite balanced. Nonetheless, all three potential covariates were included in the final Cox proportional hazards model as potential confounders of the association between instrument and decline in physical functioning/well-being. The hazard ratio for instrument group did not change substantially (adjusted HR=0.615) from the unadjusted HR (0.623) indicating an absence of confounding.

Table 4.14: Associations between instrument and potential co-variates with rate of HRQL decline in Physical Functioning/Well-being

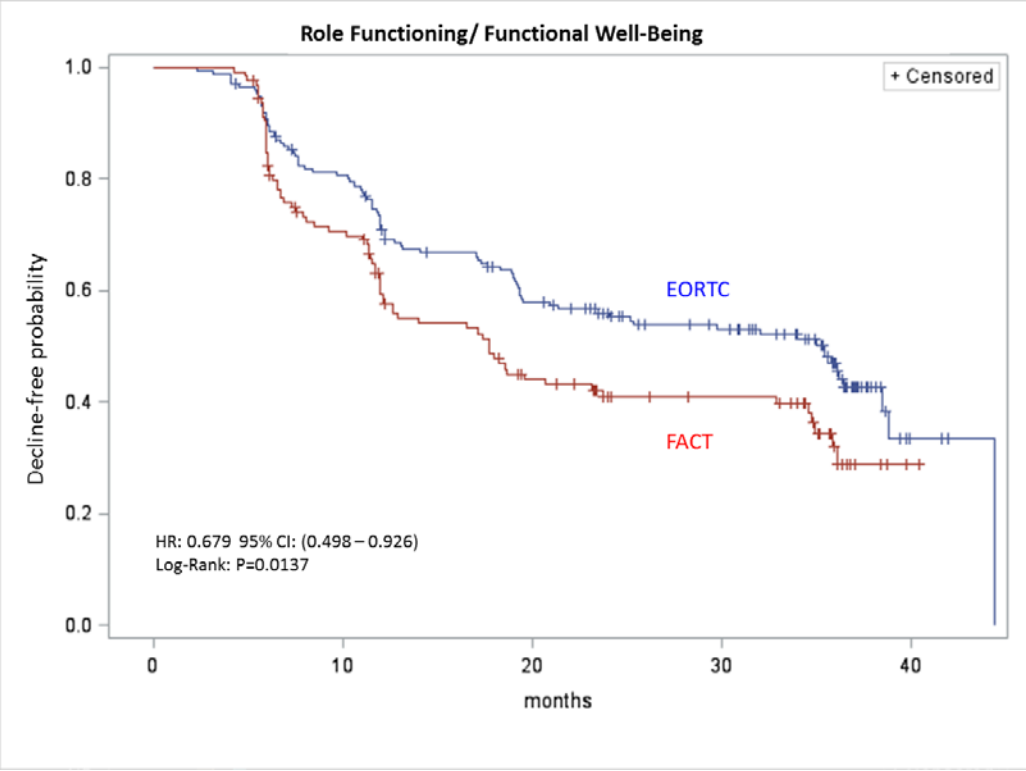
Variable	Unadjusted model			Adjusted model		
	HR	95% CI	p-value	HR	95% CI	p-value
Instrument						
FACT	Reference					
EORTC	0.623	0.456-0.851	0.0029	0.615	0.449-0.844	0.0026
PSA						
>50	Reference					
20-50	0.788	0.535-1.159	0.2255	0.809	0.547-1.196	0.2877
<20	0.566	0.385-0.833	0.0038	0.564	0.382-0.834	0.0041
AGE						
>=70	Reference					
60-69	1.257	0.920-1.717	0.1515	1.227	0.893-1.685	0.2063
50-59	0.529	0.214-1.307	0.1678	0.468	0.189-1.160	0.1009
40-49	0	0	0.9749	0	0	0.9754

GLEASON SCORE						
< 8	Reference					
8 - 10	0.761	0.553-1.048	0.094	0.759	0.550-1.048	0.0936

Time-to-HRQL decline in the role functioning/functional well-being domains

Figure 4.2 shows the Kaplan-Meier curves for decline in role functioning/functional well-being domains. Similar to the curves for the physical domain, the FACT-P instrument showed a significantly greater rate of decline compared to the EORTC QLQ-C30+3 (log-rank p=0.0137) and this is apparent starting at the time of the first assessment at 6 months up until the 3 year mark. The mean time-to-decline was 27.7 months for patients who completed the EORTC QLQ-C30+3 and 21 months for patients who completed the FACT-P.

Figure 4.2: Time-to-HRQL decline for role functioning/functional well-being by instrument allocation



Crude and adjusted associations between instrument and role functioning/functional well-being

Similar to the physical functioning/well-being domain, adjusting for potential confounders did not change the strength or significance of the association between instruments and decline in role functioning/functional well-being. Covariates included (based on association with decline at p-value of <0.2) were PSA and clinical stage. Both variables however did not have statistically significant associations with decline in the unadjusted or adjusted models. The final cox proportional hazards model yielded an adjusted HR of 0.668 for Instrument (vs. unadjusted HR=0.679), thus PSA and clinical stage did not confound the relationship between Instrument and decline in role functioning/functional well-being (see table 4.15 below).

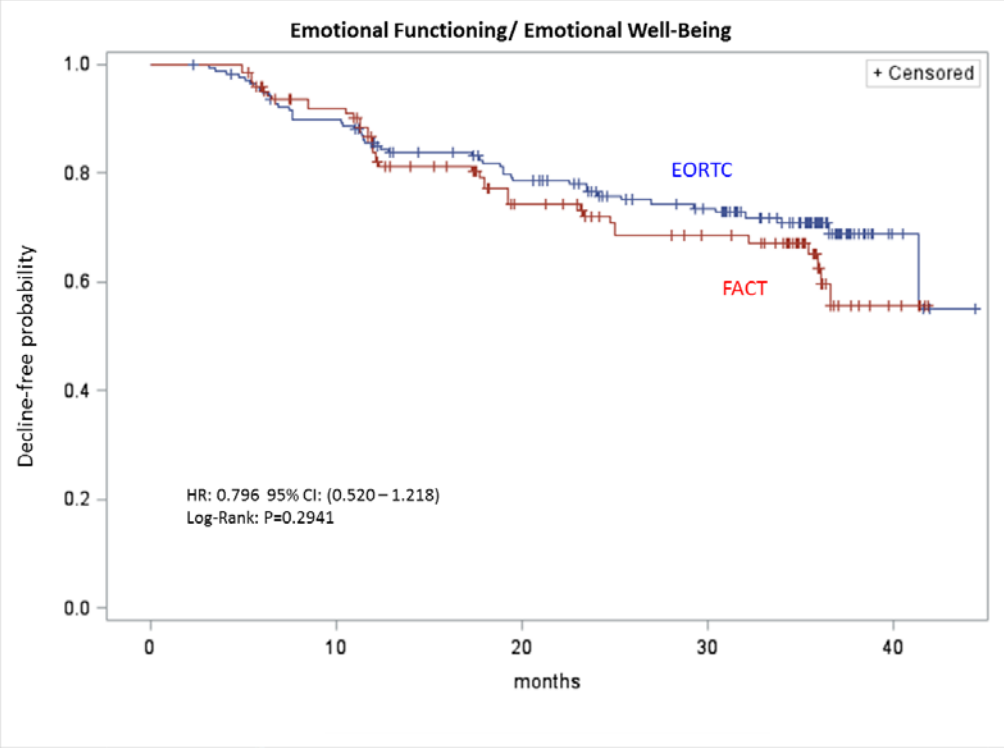
Table 4.15: Associations between instrument and potential covariates with rate of HRQL decline in Role Functioning/Functional Well-being

Variable	Unadjusted model			Adjusted model		
	HR	95% CI	p-value	HR	95% CI	p-value
Instrument						
FACT	Reference					
EORTC	0.679	0.498-0.926	0.0145	0.668	0.481-0.927	0.0159
PSA						
>50	Reference					
20-50	1.487	0.974-2.272	0.0663	1.533	0.975-2.409	0.0641
<20	1.089	0.714-1.662	0.6917	1.185	0.754-1.861	0.4612
CLINICAL STAGE						
4	Reference					
3	1.314	0.614-2.808	0.4817	1.503	0.691-3.270	0.3043
2	2.435	0.816-7.265	0.1105	2.49	0.825-7.513	0.1054

Time-to-HRQL decline in emotional functioning/emotional well-being

The Kaplan Meier curves were not significantly different between instruments for Emotional functioning/well-being as shown in figure 4.3. However, it can be observed from the curves that the FACT-P group for the most part does show greater rate of decline compared to the EORTC QLQ-C30+3, consistent with the other functional domains. The mean time-to-decline was 33.6 months for patients completing the EORTC QLQ-C30+3 and 29.4 for patients completing the FACT.

Figure 4.3: Time-to-HRQL decline for emotional functioning/emotional well-being by instrument allocation



Crude and adjusted associations between instrument and emotional functioning/ well-being

The unadjusted HR of 0.796 indicates that the EORTC QLQ-C30+3 instrument reports a lower rate of decline in emotional functioning/well-being compared to the FACT-P. In evaluating

potential confounders, only performance status was found to be associated with decline in emotional functioning/well-being at a p value <0.2. When this was factored in the model, the HR for instrument changed from 0.796 (unadjusted) to 0.734 (adjusted) however this change was less than 10% therefore performance status did not confound the relationship between instrument and emotional decline (see table 4.16).

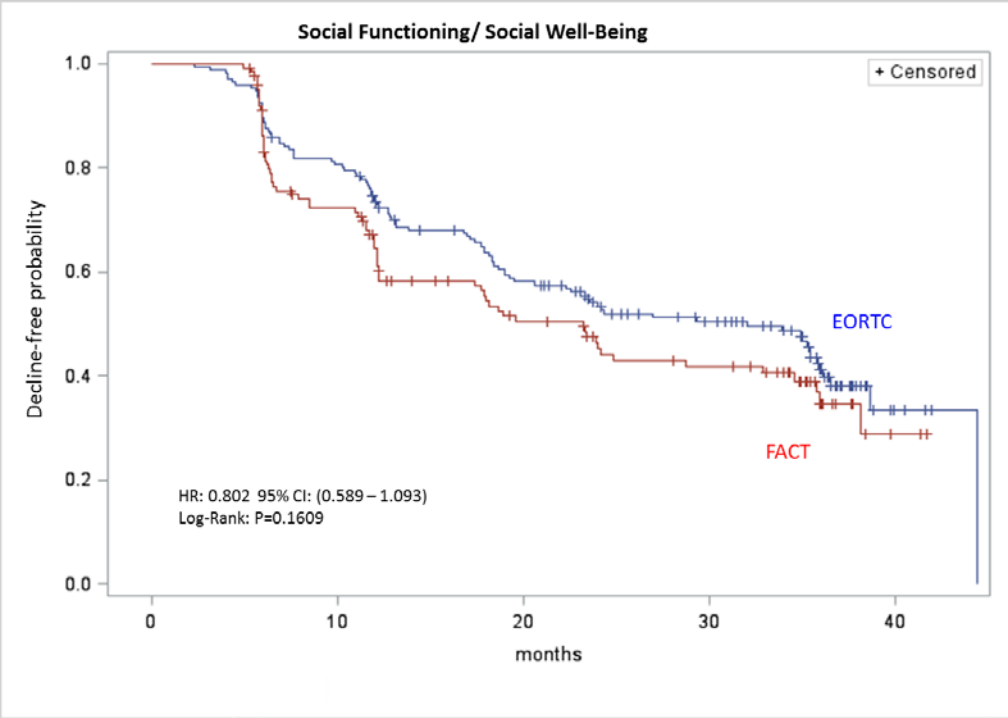
Table 4.16: Associations between instrument and potential covariates with rate of HRQL decline in emotional functioning/emotional Well-being

Variable	Unadjusted model			Adjusted model		
	HR	95% CI	p-value	HR	95% CI	p-value
INSTRUMENT						
FACT	Reference					
EORTC	0.796	0.520-1.218	0.2927	0.734	0.479-1.125	0.1562
PERFORMANCE STATUS						
0	Reference					
1/2	0.298	0.109-0.815	0.0184	0.281	0.102-0.769	0.0135

Time-to-HRQL decline in social functioning/social well-being

There was no significant difference observed in the Kaplan Meier curves for social functioning/well-being as shown in figure 4.4. However, the trend can again be observed that the FACT-P reports a greater rate of decline compared to the EORTC QLQ-C30+3. The mean time-to-decline was 27.3 months for patients completing the EORTC and 22.7 for patients completing the FACT.

Figure 4.4: Time –to-HRQL decline for social functioning/social well-being by instrument allocation



Crude and adjusted associations between instrument and social functioning/ well-being

The unadjusted HR of 0.802 indicates a lower rate of decline in social functioning/well-being is observed with the EORTC QLQ-C30+3 instrument compared to the FACT-P instrument. Potential confounders associated with HRQL decline (at p value <0.2) included, age, rectal exam, Gleason score, and PSA. When these covariates were included in the model, the HR for instrument changed from 0.802 (unadjusted) to 0.811 (adjusted) however, again, there was no confounding due to any of these variables as the difference between adjusted and unadjusted HR was less than 10% (see table 4.17).

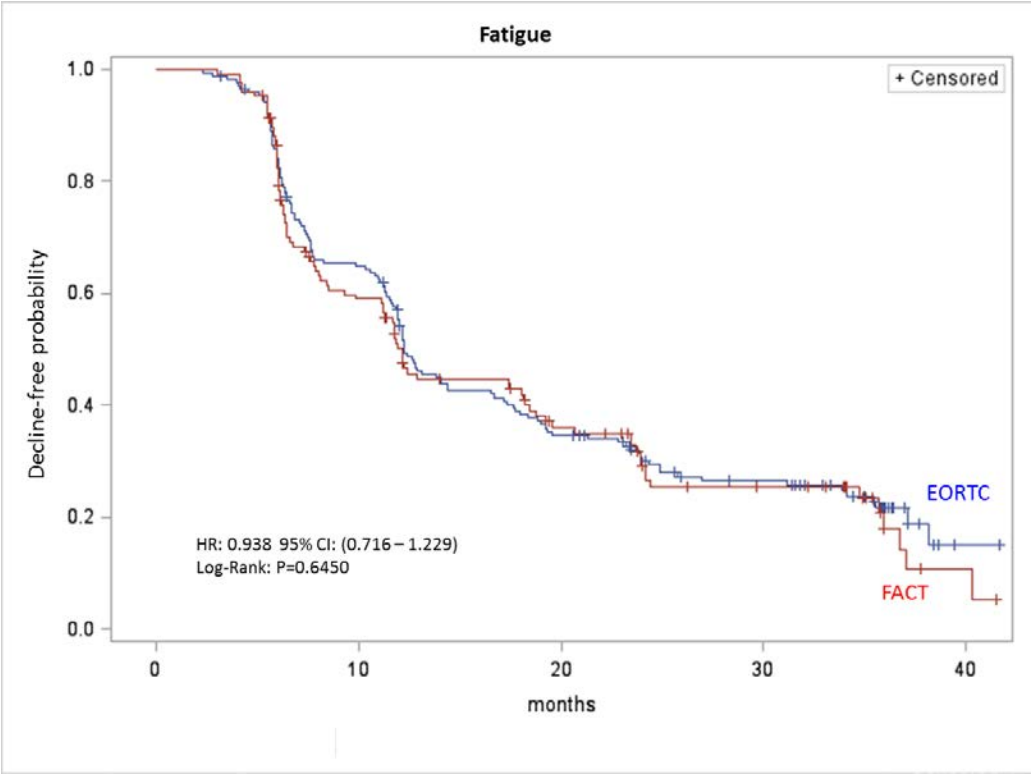
Table 4.17: Associations between instrument and potential covariates with rate of HRQL decline in social functioning/social well-being

Variable	Unadjusted model			Adjusted model		
	HR	95% CI	p-value	HR	95% CI	p-value
INSTRUMENT						
FACT	Reference					
EORTC	0.802	0.589-1.093	0.1627	0.811	0.589-1.116	0.1987
AGE						
>=70	Reference					
60-69	1.124	0.823-1.536	0.4629	1.091	0.795	1.497
50-59	0.521	0.211-1.288	0.158	0.565	0.226-1.412	0.2215
40-49	0.891	0.123-6.464	0.9092	1.57	0.184-13.422	0.6805
RECTAL EXAM						
Y	Reference					
N	1.666	0.878-3.161	0.1181	1.641	0.826-3.258	0.1573
GLEASON SCORE						
< 8	Reference					
8 - 10	0.717	0.520-0.989	0.0429	0.707	0.510-0.980	0.0377
PSA						
>50	Reference					
20-50	0.72	0.481-1.079	0.1112	0.784	0.518-1.189	0.252
<20	0.767	0.525-1.120	0.169	0.814	0.549-1.209	0.308

Time-to-decline in fatigue

The Kaplan Meier curves for fatigue displayed in figure 4.5 indicate that the 2 instruments are actually remarkably similar in terms of their ability to detect decline in fatigue. The HR of 0.938 indicates that there is very little difference between the 2 curves (p=0.6450). The mean time-to-decline was 18.2 months for patients completing the EORTC QLQ-C30+3 and 18.0 for patients completing the FACT-P.

Figure 4.5: Time–to-HRQL decline for fatigue by instrument allocation



Crude and adjusted associations between instrument and fatigue

Potential confounders included age, clinical stage, rectal exam and PSA (associated with decline in fatigue at a p value <0.2). When these variables were included in the model, the HR for instrument changed from 0.938 (unadjusted) to 0.952 (adjusted). Similar to the functional domains, this change was less than 10% therefore these factors did not confound the relationship between instruments and fatigue (see table 4.18).

Table 4.18 Associations between instrument and potential covariates with rate of HRQL decline in Fatigue

Variable	Unadjusted model			Adjusted model		
	HR	95% CI	p-value	HR	95% CI	p-value
INSTRUMENT						
FACT	Reference					
EORTC	0.938	0.716-1.229	0.6439	0.952	0.712-1.272	0.7378
AGE						
>=70	Reference					
60-69	1.328	1.013-1.742	0.0399	1.324	0.995-1.763	0.0544
50-59	0.596	0.291-1.224	0.1588	0.592	0.266-1.318	0.1991
40-49	0	0	0.9785	0	0	0.9785
CLINICAL STAGE						
4	Reference					
3	0.506	0.303-0.846	0.0094	0.551	0.321-0.946	0.0306
2	0.274	0.091-0.830	0.0221	0.238	0.077-0.734	0.0125
RECTAL EXAM						
Y	Reference					
N	1.451	0.828-2.543	0.1931	1.14	0.621-2.091	0.6733
PSA						
>50	Reference					
20-50	0.868	0.609-1.238	0.4349	0.822	0.559-1.209	0.319
<20	0.722	0.513-1.017	0.0623	0.622	0.429-0.901	0.0121

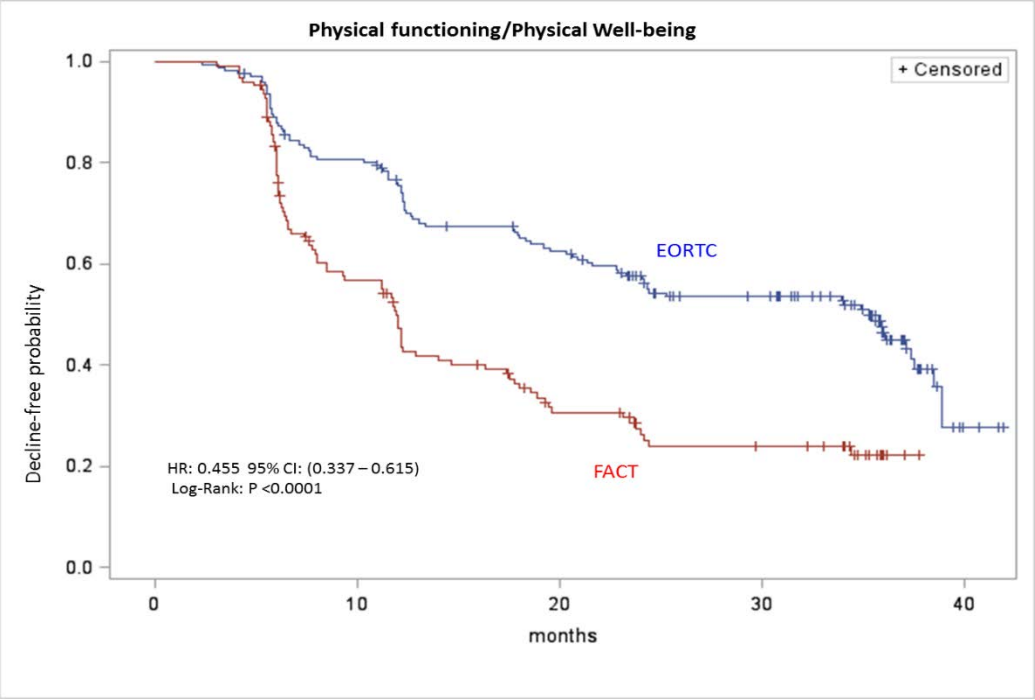
In summary, although the between-instrument group differences were only statistically significant for the physical and role/functional domains the trend was consistent for all four common functional domains. The rate of HRQL decline was greater for patients completing the FACT-P questionnaire compared to the patients who completed the EORTC QLQ-C30+3. The instruments reported a similar rate of HRQL decline, however, for fatigue.

Sensitivity analysis

In a sensitivity analyses, the time-to-HRQL decline analyses were repeated with the cut-off point for clinically meaningful decline reduced to 7% (or 7 point) from 10% (10 points). The Kaplan Meier curves are displayed below, in figures 4.6 to 4.10.

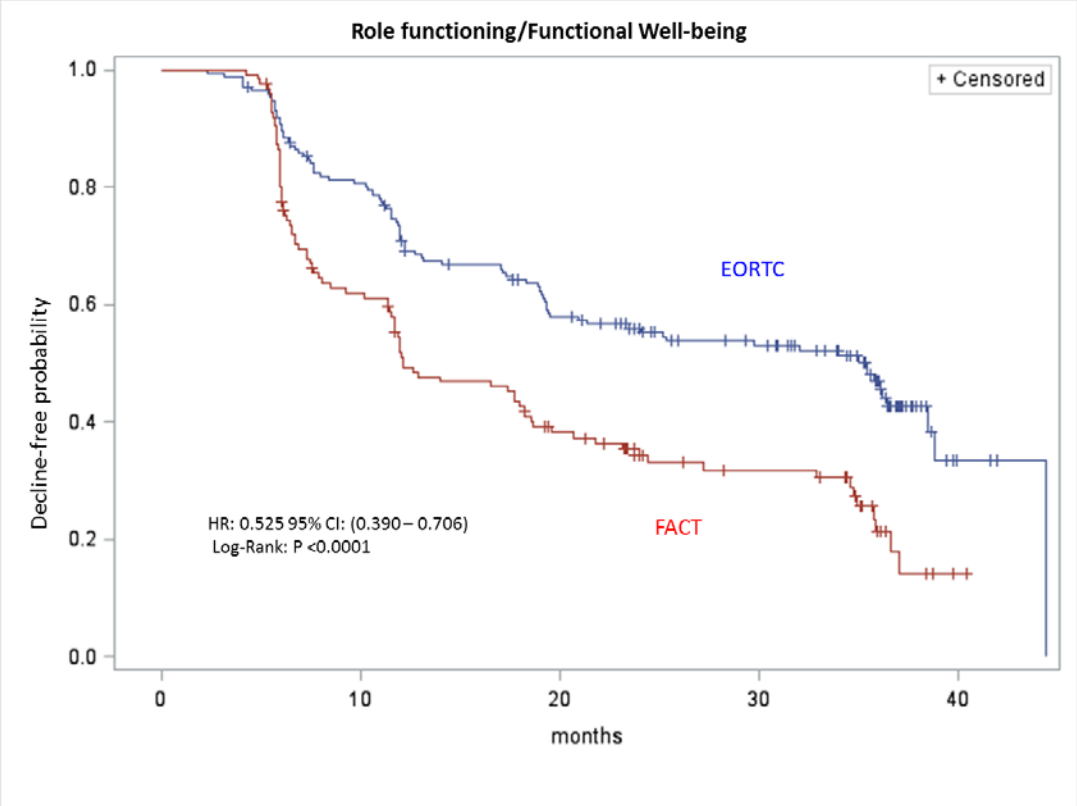
For the physical domain, reducing the clinically meaningful change cut-point resulted in a more pronounced difference between the 2 instruments as indicated by the larger separation between the Kaplan Meier curves (see figure 4.6), a stronger hazard ratio and smaller p-value. The hazard ratio was reduced from 0.623 to 0.455 and the log-rank p-value was reduced from .0026 to <0.0001. The FACT-P remained the instrument showing the greater rate of HRQL decline, compared to the EORTC QLQ-C30+3.

Figure 4.6: Time-to-HRQL decline for physical functioning/physical well-being (sensitivity analysis)



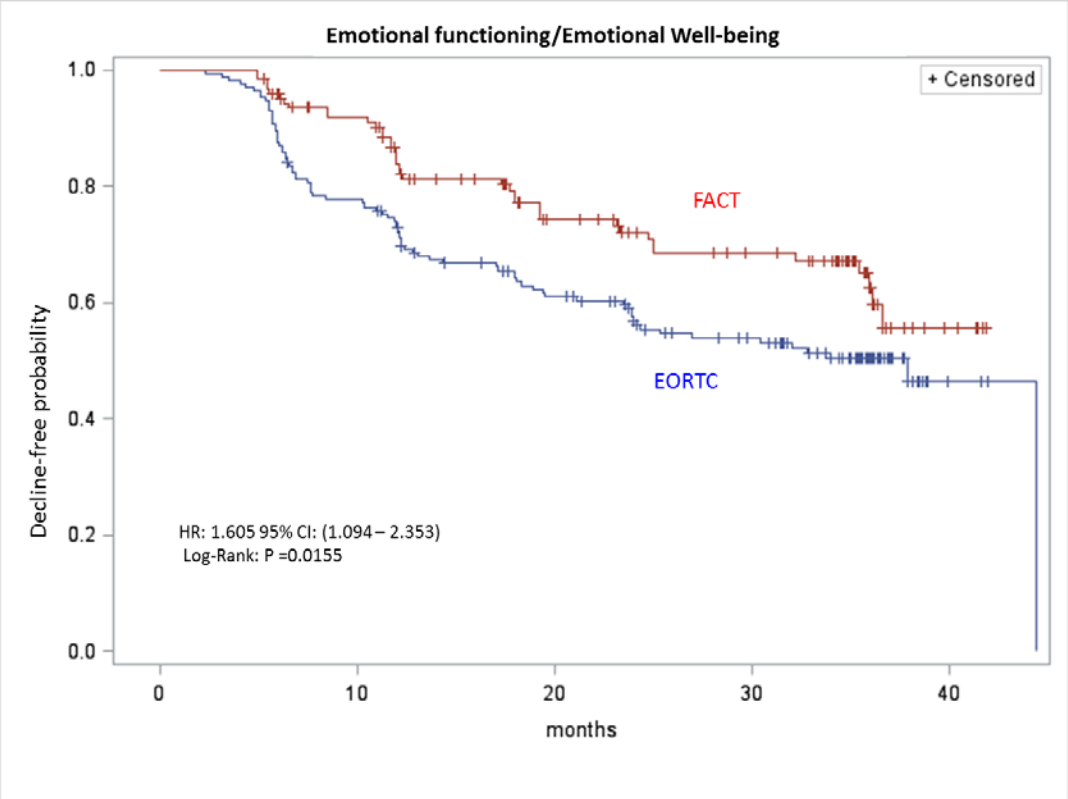
Consistent with the physical domain, larger differences were observed between instruments for the role functioning/functional well-being domains when the clinically meaningful change cut-point was reduced. Although the difference was statistically significant at the 10% cut-point as well, the hazard ratio was reduced to 0.525 (from 0.679) and the log-rank p value was reduced from 0.0137 to <0.0001 when the cut point was changed to 7%. This can be visualized from the more pronounced separation of the Kaplan Meier curves in Figure 4.7 below.

Figure 4.7: Time-to-HRQL decline for role functioning/functional well-being (sensitivity analysis)



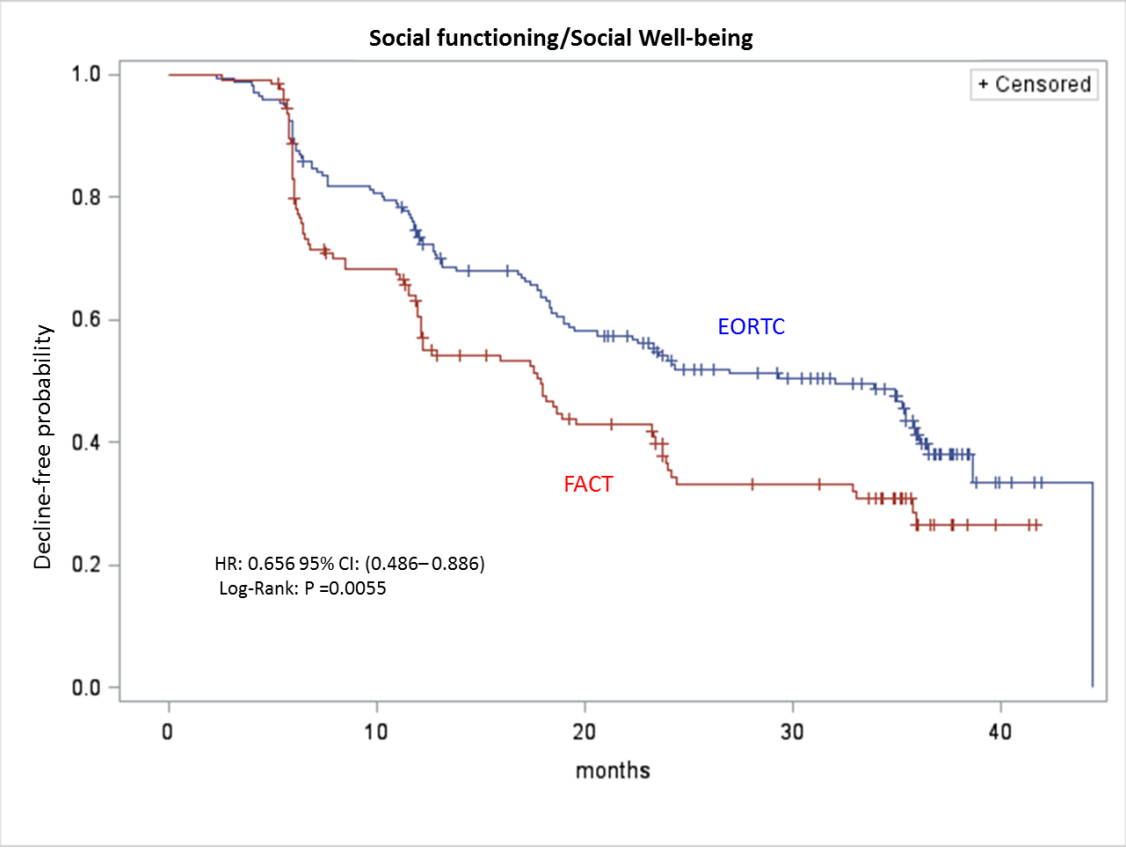
For emotional functioning/well-being, reducing the cut-point actually had a quite different effect from the other functional domains. The difference between the 2 instruments became statistically significant in terms of time-to-decline when the 7% cut-point was used for clinically meaningful change, where it was not significant at the 10% cut-point. However, the EORTC QLQ-C30+3 became the instrument with the greater rate of HRQL decline (whereas the FACT-P showed greater rate of decline at the 10% cut-point).

Figure 4.8: Time-to-HRQL decline for emotional functioning/emotional well-being (sensitivity analysis)



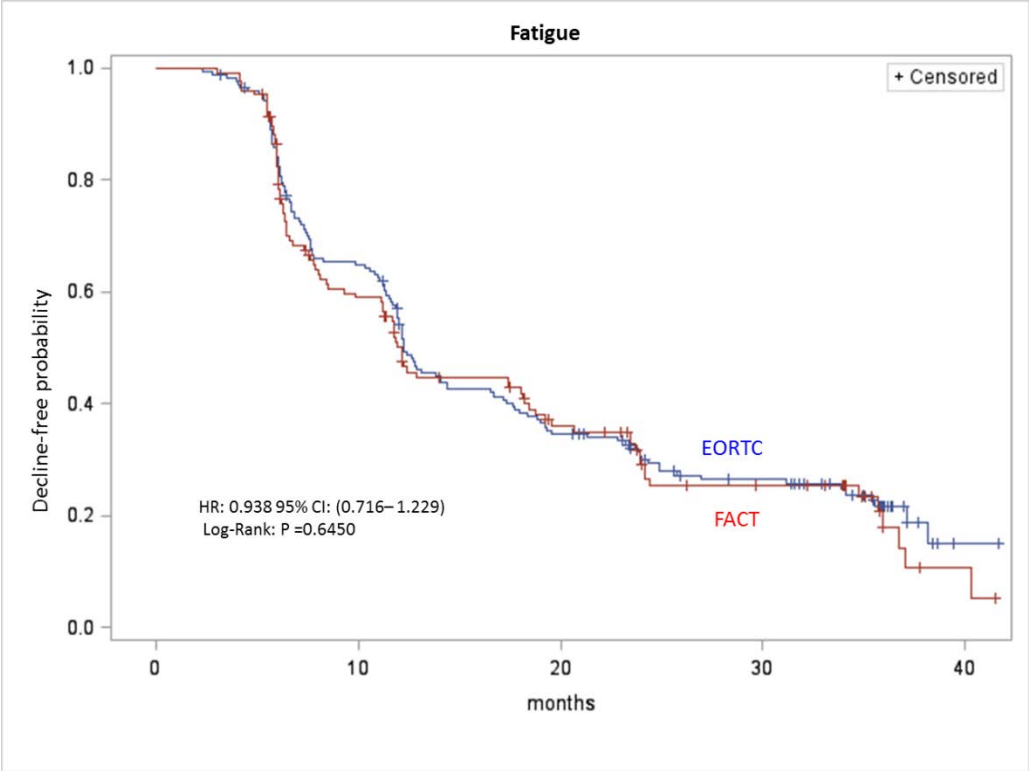
The difference in time-to-HRQL decline between instruments for the social functioning/social well-being domain was not statistically significant when a clinically meaningful difference cut-point of 10% was used. However, when the cut-point was reduced to 7%, a significant difference was observed between the 2 instruments (log-rank p=0.0055). The FACT-P showed a greater rate of decline than the EORTC QLQ-C30+3 (HR=0.656), consistent with the original analysis. Figure 4.9 below displays the Kaplain Meier curves.

Figure 4.9: Time-to-HRQL decline for social functioning/social well-being (sensitivity analysis)



Fatigue was the only item/domain analyzed that had no change in time-to-decline for both instruments when the cut-point was reduced from 10% to 7%. In fact, the hazard ratio and P-values remained exactly the same (see figure 4.10). There was no difference observed between instruments in time-to-HRQL decline for fatigue (HR=0.938, log-rank P value=0.6450).

Figure 4.10: Time-to-HRQL decline for fatigue (sensitivity analysis)



In summary, it was observed that the between-instrument group differences became more pronounced for all 4 common functional domains (physical, role/functional, emotional and social) when the cut-point for clinically meaningful change was reduced from 10% (10 points) to 7 % (7 points) in the sensitivity analyses. For the social and emotional domains the between-instrument group differences in time-to-HRQL decline became significant when the CMC cut-point was changed to 7% (whereas the differences were not significant when the 10% cut-point was used).

For the physical and role/functional domains the difference between instruments was significant at both CMC cut-off points. For 3 of the 4 domains (physical role/functional and social), the FACT-P remained the instrument with the greater rate of decline compared to the EORTC QLQ-C30+3. For emotional functioning/well-being however, the direction of the association changed in the sensitivity analysis such that the hazard ratio became >1 (indicating the EORTC QLQ-C30+3 reported a greater rate of decline compared to the FACT-P). For fatigue, the 2 instruments were similar in terms of time-to-HRQL decline at both the 10% and 7% CMC cut-point.

Chapter 5

Discussion

The purpose of this thesis project was to compare two widely used HRQL instruments in terms of the results they produce in a subset of patients from a randomized controlled trial of men with locally advanced prostate cancer and to identify the differences in responsiveness, if any, between the two instruments using a variety of HRQL analysis methods. This chapter will focus on the interpretation of these results, and will discuss the strengths and weaknesses of this project, and the contributions the findings make to current research in this area.

5.1 Summary of results and interpretations

This study utilized data from an *a priori* conceptualized sub-study incorporated into the PR.3 study, a phase III clinical trial which aimed to investigate the benefit of adding radiation therapy to hormonal therapy in men with locally advanced prostate cancer. HRQL was an important secondary endpoint of this study, as it is well known that radiation therapy (RT) is associated with toxic effects in addition to decline in some symptomatic aspects of HRQL. Therefore, it was important for researchers, clinicians and future patients to know whether the benefits of RT (if found to have an overall survival advantage) were offset by detrimental effects on HRQL. RT is associated with sexual dysfunction, bowel and bladder irritation, and bowel obstructive symptoms, and has been shown to negatively impact these corresponding HRQL symptoms scores as well (21, 22). Hormonal therapy (ADT) is also associated with adverse effects including sexual dysfunction, fatigue, anemia, loss of bone density, muscle atrophy and alterations in myocyte contractions and has been found to have a negative impact on HRQL physical function scores (20). As this thesis project is concerned primarily with the comparison of HRQL instruments (and to a lesser extent the impact of RT on HRQL), the impact of ADT on HRQL is of particular interest. The 2 instruments we sought out to compare measure four of the

same functional domains, one of which is physical functioning. The functional scales are included in the core aspects of the both questionnaires, which are widely used in clinical trials of all cancer and treatment types. Therefore, the results of the comparisons of the functional domain scores were of particular interest in this study. The symptom domains (relevant to RT) on the other hand, while of the most interest to the PR.3 study, were of secondary interest in this project. The symptom scores were generated from the PR17 checklist (administered with the EORTC QLQ-C30+3) and the prostate cancer subscale (on the FACT-P), and while the FACT-P is widely used and has been well validated the PR17 checklist was a list of questions put together specifically for the PR.3 study. Thus the PR17 results, while informative to the main PR.3 study in terms of RT effects, are not likely to inform future decisions regarding the use of a prostate module (due to the availability of the now validated EORTC QLQ-PR25 module). Therefore, while evaluating each instrument in terms of ability to detect differences between treatment groups for symptom scores (Objective 2) was an interesting and informative aspect of this project; the main focus was to compare the change in functional scores between the 2 instruments (Objective 3).

5.1.1 Objective 2: Between-treatment arm comparison (ADT vs ADT+RT) by instrument

Comparing the 2 treatment arms (ADT vs ADT+RT) by instrument was intended to identify if there are differences in the ability to detect changes in HRQL symptom scores, attributable to radiation, between the two instruments. Thus, for this objective we compared the instruments in terms of their respective *discriminative* performance. The results of our between-treatment arm comparison for symptoms scores at 6 months indicated that the FACT-P was able to detect a statistically significant between-treatment arm difference in mean change scores for the urine symptom domain whereas the PR17 (administered with the EORTC QLQ-C30+3) was not able to show a significant between-treatment arm difference. A comparison of proportions of patients worsened, improved and stable (where worsened and improved included patients that met the clinically meaningful change threshold of 10% decrease or increase respectively, from

baseline) did not indicate significant between treatment arm differences for either instrument. Thus in terms of the urine symptom domain scores, we conclude that the FACT-P is more responsive to detecting between-treatment arm differences (i.e. to detecting the impact of RT on urine symptoms). The EORTC QLQ-C30+3/PR17 was clearly superior, however, in terms of detecting between-treatment arm differences in bowel/rectum scores as well as diarrhea. Between-treatment arm differences for both mean change scores and for proportions of patients improved, stable and worsened were statistically significant for the EORTC QLQ C-30/PR17 bowel/rectum symptoms and diarrhea scores but not for the FACT-P bowel trouble score. The FACT-P unfortunately is lacking in terms of capturing bowel/rectum symptoms and diarrhea as it contains only 1 item related to bowel, and this item ("Difficulty moving one's bowels") is not particularly relevant to radiation effects. It is important to note, however, that the comparisons done in objective 2 are purely descriptive and therefore no assumptions can be made as to whether the two instruments are statistically different in terms of detecting differences between treatment groups. We elected not to conduct a statistical comparison of mean change scores between instruments, because the 2 instruments have different underlying scales (i.e. a change score on one instrument is not necessarily equivalent to a change score on the other). Further, we were not confident that comparing a common linear transformation of the scales would be valid.

5.1.2 Objective 3: Between-instrument comparison (EORTC QLQ-C30+3/PR17 vs FACT-P)

For Objective 3 we used a number of methods to compare the 2 instruments directly with respect to their ability to detect change in HRQL over time, for both symptom scores and functional domains scores. The analyses performed were intended to investigate if the instruments are different in terms of the results they produce, specifically in detecting changes in HRQL attributable to hormonal therapy (which all participants were to receive on study) and also in terms of detecting short term changes attributable to radiotherapy (which approximately half of

the patients in each instrument group were to receive). Thus, the aim of this objective was to compare the 2 instruments in terms of their *evaluative* performance.

Although we had already investigated between-treatment arm differences in symptom scores, we also thought would be interesting to look at between-instrument differences in proportions of patients worsened, improved and stable for symptoms scores at 6 months both with treatment arms combined and in the ADT+RT arm only. Statistically significant differences were observed between instruments for the urine domain scores, with the FACT-P reporting a greater proportion of patients with worsened urine scores compared to the PR17 (when treatment arms were combined). Thus, our conclusions were in line with that of objective 2, that the FACT-P is more responsive to clinically meaningful worsening in urine symptoms attributed to radiotherapy.

For functional domain and fatigue scores, no statistically significant differences were observed between the 2 instruments in terms of proportions of patients improved/stable and worsened for any of the 4 common functional domains and for fatigue. This was the case for both the analysis at 36 months, and the cumulative analysis (up to and including 36 months). However, there was a consistent trend observed in the cumulative analysis in that for all 4 functional domains (physical, social, emotional and role/functional) the FACT-P showed a slightly greater proportion of patients were worsened at some point from baseline to 36 months compared to the EORTC QLQ-C30+3. When we reduced the clinically meaningful change cut-point to 7% (from 10%) in the sensitivity analysis, the between instrument differences became significant for 3 of the 4 functional domains (physical, emotional and role/functional), with both physical and role/functional domains in favour of the FACT-P showing a greater proportion of worsened patients. Therefore, these results suggest that the FACT-P may be more responsive to detecting clinically meaningful changes in functional domains over time. Fatigue measurement was robust in both instruments, as there was no statistically significant difference between instruments at either CMC cut-point.

Statistically significant differences in time to detectable clinically meaningful HRQL decline were observed between the 2 instruments for the both the physical and the role/functional domain scores. Between instrument differences were also observed for the social and emotional domains, although these differences were not statistically significant. For all four common functional domains, FACT-P reached the threshold for clinically meaningful HRQL decline sooner that the EORTC QLQ-C30+3.

We again performed a sensitivity analysis, altering the clinically meaningful change cut-point from 10% to 7%. The purpose of this analysis was to test the robustness of our results by altering the criteria for a clinically meaningful change (CMC), given that the literature supports a 5-10% range for CMC (33, 39, 42). Reducing the CMC threshold resulted in a stronger association between instrument-specific detection of a clinically meaningful decline, and between-instrument group differences were statistically significant for all 4 common functional domains. Once again, the FACT-P showed a significantly greater rate of decline for 3 of the 4 functional domains (physical, role/functional, and social). For the emotional domain, however, the reduction of the CMC cut-point to 7% resulted in a reversal of the association between HRQL decline and instrument. The EORTC QLQ-C30+3 reported a significantly greater rate of decline in emotional functioning compared to the FACT-P. Similar to the response analysis, the FACT-P and EORTC QLQ C-30+3 were similar in terms of measuring time to HRQL decline for fatigue, at both CMC cut-points.

Based on our observations using the methods we did in this study, we found that the FACT-P was more responsive to HRQL changes attributed to long term ADT therapy compared to the EORTC QLQ-C30+3. There are some important distinctions between our study and other studies comparing the responsiveness of the EORTC QLQ-C30 and the FACT that make it somewhat challenging to put our study in context of other published work. No previous studies have directly randomized patients to an HRQL assessment. However, there are certainly some consistencies

observed between our results and other studies. Our findings were similar to those of Conroy et al, who compared individual subscales within the EORTC QLQ-C30 and FACT-G and found that the FACT Physical and Functional Well-Being were amoung the most responsive subscales (49). Our findings were not consistent with Uwer et al., who found that the EORTC QLQ-C30 was the more responsive instrument. However Uwer's study was in colorectal patients receiving either chemotherapy or radiation, and the EORTC QLQ-C30 was only found to be responsiveness in patient's receiving chemotherapy (30). Our study on the other hand, was mainly concerned with looking for changes in HRQL associated with long term hormonal treatment in men with prostate cancer. Secondly, responsiveness was evaluated in the Uwer et al. study using standardized response mean (SRM) and Effect size (ES), where SRM was the mean changes in scores between baseline and follow-up divided by the standard deviation (SD) of this change; and the ES was the mean change in scores between baseline and follow-up divided by the SD of the baseline score. Responsiveness was assessed using data from patients deemed to have improved, as determined by the patients overall assessment of his/her change in health status. On the contrary, our study was concerned with deterioration of HRQL. The study by King et al. is the only earlier study that has compared the 2 instruments directly with statistical methods (54). The researchers in this study essentially used a distribution-based approach to calculate a responsiveness index (RI), which was the mean change in the intervention arm divided by the SD of change in the control arm. Their study however was a secondary analysis which utilized data from a study designed to evaluate the impact of an intervention on QOL, where the researchers were expecting to observe improvement in the intervention group. Contrary to the King et al. study this thesis project addressed detecting deterioration in HRQL, due to the nature of the intervention (ADT treatment). We also evaluated responsiveness in a different way for the functional domains. Nonetheless, there were some similarities with respect to the results of the King et al. study and our study. Although the differences were not statistically significant, King et al. did find that the

FACT was more responsive than the EORTC for Physical and Role/Functional domain which was consistent with our findings. The response analysis (comparison of % improved/stable vs worsened) and time to clinically meaningful HRQL decline analysis is perhaps another approach to compare responsiveness of 2 instruments, and one we were able to utilize in our study due to the randomized sub-study design for allocation of instruments.

5.2 Methodological issues and limitations

There are some limitations and potential methodological issues in this study that must be taken into consideration when making conclusions about the results observed. These limitations and methodological considerations will be discussed in this section.

5.2.1 Potential sources of bias and confounding

It is important to emphasize that *individual* patients were not randomized to an instrument (as they were to treatment) but rather the *centre* that participated in the study was randomly allocated an instrument (to be completed by all patients registered to the study at that centre). While the intent was to create 2 equal groups in terms of number as well as distribution of treatment arm allocation and other factors, a cluster randomization such as this does have some disadvantages. In the case of this study, the numbers of patients in each instrument group were not balanced despite the stratification by expected accrual size (179 on the EORTC QLQ-C30+3/PR.17 group and 132 in the FACT-P group). Also, with a cluster randomized designed, individuals in the same cluster (in this case, cancer centre) may be correlated or non-independent (55). In other words, 2 patients from the same centre are more likely to be similar than 2 patients from different centres (55). Cluster randomization designs also have the potential, to introduce a form of selection bias known as recruitment bias. This can occur when patients have foreknowledge of their allocation and that knowledge may impact their decision to participate in the trial (56). This would not have occurred in our case, as allocation to a HRQL instrument would not impact a participants decision to participate (the main randomization was to treatment arm, not instrument) and patients

were not aware of their centre's instrument allocation. Furthermore, although we did not perform any within group correlation calculations, we do know that the 2 instrument groups were quite similar with respect distribution of treatment allocation and baseline factors, so we feel the cluster randomization design was unlikely to have been a methodological concern.

Loss to follow up, while it can be a potential source of selection bias in clinical trials, was not of particular concern in this study as we know that HRQL questionnaire compliance was good (>85%) until 3 years post randomization to the PR.3 study and was similar between the 2 instrument groups. Differences in duration of treatment between instrument groups could be a potential source of bias if, for example, patients had a shorter duration of hormonal therapy (on average) in the one of the instrument groups compared to the other. The HRQL scores might appear higher in one group because fewer patients were being treated long term and therefore fewer would suffer long term negative HRQL effects from hormonal treatment. However, instrument allocation is very unlikely to be associated with early treatment discontinuation (other than by chance) and therefore it is unlikely that bias would be introduced this way. This is more of a concern when comparing two treatment groups, as type of treatment (associated with side effects) can be associated with early discontinuation. Since treatment allocation was balanced between the 2 instrument groups, this was unlikely to be an issue in this study.

In the time to clinically meaningful HRQL decline analysis, potential confounders were evaluated by performing a simple univariate cox proportional hazards model for ten baseline factors, including study treatment allocation, to determine if any of these factors are associated with decline in HRQL. Those associated with HRQL decline (p value ≤ 0.2) were included in the final model. The rationale for doing this was the fact that it was unknown which variables may be associated with decline in each of the domains. Due to the randomized design however, most baseline factors were relatively balanced in the 2 instrument groups. Furthermore, none of these factors were likely to truly confound the relationship between instrument and decline in HRQL,

as they were unlikely to be associated with the instrument allocation (except for by chance). There is also no biological explanation for some of these baseline factors to be associated with decline in HRQL. Although there were a number of factors that were associated with HRQL decline in the various functional domains (at a p value ≤ 0.2), including these factors in the final cox proportional hazards model did not change the HR more that 10%, indicating there was unlikely to be any confounding. This is not unexpected, as the instrument groups were relatively balanced with respect to these baseline factors.

One potential concern in this study is missing data, which is an inherent problem in HRQL measurement both in terms of missing questionnaires and missing items within completed questionnaires. Questionnaires with missing items were included in this analysis and the standard approach was used to handle missing items, that is, to generate a domain score if at least half of the items that make up that domain score are answered. If more than half of the items comprising a score were unanswered, the score would be considered 'missing' (i.e. that patient would not be included in the denominator for that domain). A thorough assessment of the missing data was performed as part of objective 1. Both missing items and questionnaire compliance (proportion of questionnaires submitted from those that were expected) were compared between the 2 instrument groups, as well as a combined comparison of both. Compliance was quite similar in both instrument groups at assessment time points from baseline to 3 years, which is reassuring. An assessment of the overall missing data (factoring in missing items and missing questionnaires) indicated that the overall missing data was also similar between instrument groups at 6 months (ranged from 9% to 10.8% for the EORTC QLQ-C30+3/PR.17 and 7.5% to 9.2% for the FACT-P) however at 36 months there was more missing data in the FACT-P group (ranged from 6.6% to 13.2% in the EORTC QLQ-C30+3/PR.17 group and 18.5 to 19.6% in the FACT-P group). This missing data was across domains. If there was any potential for bias to be introduced in this way, it could be assumed that those who decided not to complete the questionnaires/items are those

that were less well and perhaps those that had worse HRQL. Therefore, if anything, the FACT-P scores may have been slightly inflated (i.e. skewed in favor of showing a greater proportion of patients stable or improved) if there was bias introduced. Since the FACT-P for the most part showed more worsening in the various analyses, there should not be much concern in terms of our results (i.e. we may have underestimated the differences between the instruments groups, if anything).

5.2.2 Study power

We had a fixed sample size of 311 patients to work with in this study, which included the North American patients accrued to the PR.3 study who completed the baseline HRQL questionnaire as per their site's allocation. There was no pre-determined sample size for this sub-study, but rather it was designed to include all patients randomized to the PR.3 study from North America up until the study was closed to accrual (when the total PR.3 sample size was reached). The North American centres were randomly allocated to one of the two instruments, and were stratified by expected accrual size. The final sample included 179 patients from sites allocated to the EORTC QLQ-C30+3/PR.17 and 132 patients from sites allocated to the FACT-P. Therefore, *a priori*, we could estimate the minimal detectable difference in proportions between instrument groups at a set sample size of 311, with power of 80% and type 1 error of 0.05 (see Appendix C**).** For objectives 1 and 2, we did not perform any *a priori* power calculations as these were descriptive comparisons between the 2 instrument groups. However, based on the results of objective 2 we know that we did have adequate power to detect differences in mean change scores and proportions improved, stable and worsened between *treatment* groups, for certain domains within both instrument groups. For objectives 3A and 3B**,** we determined *a priori* that we would have 80% power to detect a 15-16% difference in proportions of patients with a clinically meaningful decline in HRQL assuming 50% 'worsened' in one of the 2 instrument groups. In this study, there is no 'control' group as there would be for a treatment group

comparison, and we did not have any knowledge regarding which of the instruments would show greater proportion of patients with 'worsened' HRQL. We were in fact able to detect some differences between instruments for some symptoms at 6 months; however, we did not have enough power to detect any differences between instruments for functional domains and fatigue up to 36 months (see Appendix C). We would have, in fact, needed a very large sample size to detect differences as small as we observed in these analyses (see Appendix C). In a sensitivity analysis however, which adjusted the cut-point for clinically meaningful change from 10% to 7%, we were able to detect significant differences between the instrument groups for some functional domains. For objective 3C, we did not have information on the rate of clinically meaningful worsening of HRQL in men receiving ADT for early prostate cancer *a priori*, and therefore it would have been difficult to estimate the detectable effect estimate (hazard ratio). However, the statistical test for objective 3C is considered to be more efficient than that of 3A/3B, and therefore we were less concerned about power for this analysis. This was proven to be true, as we observed significant differences in time to clinically meaningful decline between the 2 instruments for the physical and role/functional domains. We did not have adequate power to detect significant differences in time to HRQL decline between instrument groups for emotional and social domains as well as fatigue. Again, we would have needed a much larger study to detect these differences (See Appendix C).

5.2.3 Other considerations

There are some notable attributes of the 2 instruments that differ and may have contributed to the differences in results observed in this study. One of these differences is in the number of response categories for the items included in the 2 questionnaires. The cut-point we used for clinically meaningful change (CMC) was in line with the literature which suggests 5-10% of the scale breadth (30, 36, 39). We used a conservative cut-point of 10% and performed a sensitivity analysis reducing the cut-point to 7%. To get a sense of what this means to a patient, it is

necessary to put it in the context of response categories. The 2 instruments differ somewhat with respect to the number of response categories for each item and this is summarized in Appendix D. For the FACT-P, there are 5 response categories (0 to 4) therefore a change in category (for example, from 0 'not at all' to 1 'a little bit') represents a 25% change in the scale. This is the case for all domain scores in the FACT-P. On the EORTC QLQ C-30+3, for 3 of the 4 functional domains common to the FACT-P (functional, social and emotional), there are only 4 response categories (1 to 4). A change of one category (for example, from 1 'not at all' to 2 'a little') represents a 33% percent change. Thus, a one category change on the EORTC QLQ-C30+3 represents a greater percentage increase in the scale compared to the FACT-P. A category change for the patient has the same meaning on either instrument; therefore it may be argued that the EORTC QLQ-C30+3 has greater potential for meeting clinically meaningful cut-point by our definition of CMC. For single items, a one category change is well above the CMC cut point for either instrument (25% or 33%) so the number of response categories would not matter. But, for domains containing multiple items, the responses for each item are averaged to generate the score. Therefore, consider a domain that is made up of 3 items. If a patient reports no change for 2 of the items and a one category change for 1 item, this would represent an overall average change of 11% for the EORTC QLQ-C30+3 and only 8.3% for the FACT-P. Thus, the change would be considered 'clinically meaningful' for the EORTC and not for the FACT, even though to the patient it is the same perceived amount of change. For the physical domain, the differences in response categories between instruments are even more pronounced. The EORTC QLQ-C30+3 has only 2 response items ('Yes' or 'No'), vs the FACT-P which as 5 (see Appendix D). Thus, even more so for the physical domain, a one category change has a greater impact on the EORTC QLQ-C30+3 (compared to the FACT-P). In our study we found that the FACT-P was more responsive in that it did show a greater proportion of patients 'worsened' for 3 common domains (at the 7% CMC cut-point only) and a greater rate of clinically meaningful decline (statistically

significant for the physical and role/functional domains, at the 10% CMC cut-point). Thus, we might conclude that the number of response categories is less important than how the two instruments contrast with respect to the way each domain is constructed.

The construction of each comparable domain differs substantially between the two instruments, both in terms of the number of items that make up a domain score as well as the types of questions. This is true for the common functional domains as well as the symptom domains/items. Appendix D compares the two instruments in terms of the number of items that make up each domain or item score. Classic test theory predicts that scales comprised of a greater number of items should be more reliable and therefore more sensitive and responsive (41). We observed the biggest differences in responsiveness between the 2 instruments were in the physical and functional/role domain scores, where the FACT-P was more responsive. The physical domain is comprised of 7 items on both instruments. As previously noted, the EORTC QLQ-C30+3 has only 2 response categories which, we would expect would make the instrument more responsive. However since this is not what we observed we conclude that the questions themselves are responsible for the differences observed. The questions are almost all uniquely different between the 2 instruments (see table 3.1 and Appendices A and B), with the exception of 1 question which is somewhat similar between the 2 instruments. For this particular population, the FACT-P therefore must ask the questions most relevant to physical HRQL changes in men with prostate cancer receiving hormonal therapy. Upon comparison of the role/functional domains in the 2 instruments, we note that the EORTC QLQ-C30+3's role domain has 2 items vs. the FACT-P's Functional domain which has 7. Therefore, the greater number of items in the FACT-P functional well-being domain may well be a factor contributing to the greater responsiveness compared with the EORTC QLQ-C30+3 role functioning domain.

Another noteworthy consideration, when discussing the limitations of measuring symptom scores, is the fact that the PR17 checklist (administered with the EORTC QLQ C30+3) was not a

validated module for assessing prostate specific symptoms. The FACT-P, on the other hand, contained the prostate specific subscale and was a fully validated instrument. Therefore in terms of measuring prostate specific symptom scores this was a limitation of the PR17. The fact that the FACT-P was more responsive to detecting urine symptoms associated with RT gives some weight to the importance of the validation process, one might conclude.

5.2.4 Generalizability

It must be acknowledged that both the EORTC QLQ-C30 and FACT questionnaires are used in many different cancer populations (in patients with many different types of cancer and for evaluating many different types of treatment). The functional domains are embedded within the core components of both instruments, therefore these domains are measured the same way for patients with prostate cancer as they are for patients with lung cancer, for example. However, the results observed in this study may only be generalizable to patients with prostate cancer who are receiving hormonal therapy, with or without radiation. We cannot assume that the same results would be observed in a different cancer population and/or in a trial assessing different treatments, as there is evidence that an instrument's responsiveness differs depending on the treatment and disease type (27). A similar type of study would need to be conducted in other cancer populations to assess responsiveness across other disease sites and treatments. Furthermore, the population in this study included patients from North America only (Canada and the US) when in fact the majority of patients randomized to the PR.3 study were from the UK. Therefore, assuming there may be some differences between the UK and North American populations, we should be cautious in generalizing these results to the entire PR.3 study population.

As previously mentioned, the PR17 checklist (administered with the EORTC QLQ C-30+3) was created specifically for the PR.3 study and was not a validated module by the EORTC. Subsequent to the PR.3 study, the EORTC QLQ-PR25 module was developed to assess prostate specific symptoms. Upon comparison of the PR17 and the EORTC QLQ-PR25 it is evident that

there are some similarities, but also a great deal of differences in terms of the composition of urine and bowel domains which were of most interest in our study (See Appendix E). For the urine domain, the EORTC QLQ-PR25 module contains 9 questions whereas the PR17 contained 8. Out of these, there were 5 questions that were similar between the two however the wording was not exactly the same for any of them. The remaining questions were unique to each of the modules. In terms of the bowel domain, the EORTC QLQ-PR25 contains 4 items vs 2 items on the PR17. Only one of these items is common between the 2 modules. Therefore, we cannot state with any certainty that the results found using the PR17 would be comparable to the results if the EORTC QLQ-PR25 was used (i.e. we cannot generalize the PR17 results to the use of the EORTC QLQ-PR25). We have shown in this study that the construction of a domain indeed makes a difference. However, we do know that the EORTC QLQ-PR25 adequately covers bowel symptoms therefore this does give it an edge over the FACT-P in terms of assessing symptoms attributed to prostate cancer radiotherapy.

5.3 Strengths and significance of findings

The design of this sub-study was quite novel, in that it is the first designed to compare two groups of patients that were randomly allocated a HRQL instrument in a clinical trial. There is not much data currently available comparing the EORTC QLQ-C30 and the FACT instruments in terms of their responsiveness. Furthermore, no studies have compared responsiveness of 2 HRQL instruments in prostate cancer and no studies have attempted to compare the two HRQL instruments in the way we have done in this project. The fact that this was a randomized sub-study (by instrument allocation) was an advantage of this study over others. We were able to compare 2 groups of patients randomly allocated to an instrument that were similar in terms of their baseline characteristics and treatments received. Therefore, the differences observed between the 2 instruments were likely due to the attributes of the instruments themselves and not to other factors. This study compared 2 of the most common instruments used to measure HRQL

in cancer. The core aspects of both questionnaires (the EORTC QLQ-C30 and the FACT-G), which contain the four common functional domains, are used across many types of cancer and cancer treatments. These results add to the current knowledge for each of these core questionnaires, and may impact the choice of researchers to use one instrument over the other in a given setting. The results of the symptom specific analyses also may be informative to researchers, if deciding whether to use the FACT-P prostate specific module or the EORTC QLQ-C30 with the current PR25 module.

5.4 Conclusions and future directions

In this sub-study of prostate cancer patients randomized to complete 2 different HRQL instruments, (the EORTC QLQ-C30+3 with PR.17 trial specific checklist or the FACT-P) there were some significant differences observed in terms of the results that each instrument provided, for some of the 'common' domains and symptoms measured. Although this study does not necessarily suggest that one instrument should be used over another, it does show that the two instruments are different despite the fact that they are intended to measure many of the same constructs. We can conclude that for this population of patients, the FACT-P was more responsive in terms of detecting 'worsening' for the most part compared to the EORTC QLQ-C30+3/PR.17. Whether the FACT-P reflects more accurately the patient experience however, we unfortunately cannot conclude from the data we had available in this study. It must be acknowledged that there is no way of knowing the 'truth' so to speak in terms of which instrument provided more accurate results, as we do not have any 'anchors' with which to correlate with HRQL findings. This is a common problem in HRQL measurement.

An important conclusion that can be made from this study is that it matters how a domain is constructed. While the two instruments investigated measured 4 common domains, there was quite a bit of variation in terms of the number and type of questions that made up the domain scores between the two instruments. This study also demonstrated the importance of selecting the

right cut-point for clinically meaningful change. We observed that changing the CMC cut-point by a few percent can quite drastically impact the results in a response analyses (comparison of % improved/stable and worsened) and for time to clinically meaningful HRQL endpoints.

For a given study, the decision to use one instrument over the other should be based on the importance of certain domains and/or symptoms as they pertain to the population and treatment intervention. Clinical trial protocols should have clear and well-articulated HRQL hypotheses and identify the relevant domains of interest (57) even when they are secondary outcomes. HRQL instruments should be reviewed to ensure that the domains and psychometric properties are sufficiently established to evaluate the HRQL hypotheses on the study, and this evidence should be provided or cited in the study protocol (57). In our study the FACT-P was more sensitive to detecting long term worsening of HRQL for functional domains. Therefore one could argue that it would be the better choice for a similar study of prostate cancer patients receiving long term hormonal therapy. However, it could also be argued that the EORTC QLQ-C30 with the current QLQ-PR25 module would be the better option for a radiation therapy study of prostate cancer patients, where the effects of treatment on diarrhea and bowel/rectum symptom scores are of great importance (since the FACT-P is limited in terms of assessing these symptoms). Further research must be done to build upon these results, and to substantiate the findings from this study.

References

1. Au HJ, Ringash J, Brundage M, Palmer M, Richardson H, Meyer RM et al. Added value of health-related quality of life measurement in cancer clinical trials: the experience of the NCIC CTG. Expert Rev. *Pharmacoeconomics Outcomes Res.* 2010 Apr(2):119-128

2. U.S Department of Health and Human Services, Food and Drug Administration, Centre for Drug Evaluation and Research (CDER), Centre for Biologics Evaluation and Research (CBER), Centre for Devices and Radiological Health (CDRH), Guidance for Industry-Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labelling Claims. Accessed March 20, 2007

3. Holzner B, Bode RK, Hahn EA et al. Equating EORTC QLQ-C30 and FACT-G scores and its use in oncological research. Eur J Cancer 2006; 42(18): 3169–3177.

4. Canadian Cancer Society's Advisory Committee on Cancer Statistics. *Canadian Cancer Statistics 2014*. Toronto, ON: Canadian Cancer Society; 2014.

5. Payne H. Management of locally advanced prostate cancer. *Asian Journal of Andrology* 2009; 11: 81–87.

6. Stratton KL and Chang SS. Locally advanced prostate cancer: the role of surgical management. *BJU International* 2009; 104, 449-454.

7. Kahn MA, Partin AW. Management of High-Risk Populations with Locally Advanced Prostate Cancer. *The Oncologist* 2003; 8:259-269.

8. Catalona WJ, Smith DS, Ratliff TL, Basier JW. Detection of Organ-Confined Prostate Cancer Is Increased Through Prostate-Specific Antigen-Based Screening. *JAMA* 1993; 270 (8):948-954.

9. Johansson JE, Holmberg L, Johansson S, Bergstrom R, Adami HO. Fifteen-Year Survival in Prostate Cancer – A Prospective, Population-Based Study in Sweden. *JAMA* 1997; 277(6):467-471.

10. Pilepech MV et al. A Phase III trial of androgen suppression before and during radiation therapy (RT) for locally advanced prostatic carcinoma, preliminary report of RTOG protocol 8610. *Proc Am Soc of Clin Oncol* 1993; 12:229.

11. Pilepech MV et al. Phase III RTOG trial 86-10 of Androgen Deprivation before and during radiotherapy in locally advanced carcinoma of the prostate. *Proc ASCO* (17) 308a, 1998.

12. Fellows GJ, Clark PB, Beynon LL, Boreham J, Keen C, Parkinson MC, Peto R, Webb JN. Treatment of Advanced Localised Prostatic Cancer by Orchiectomy, Radiotherapy, or Combined Treatment. *British Journal of Urology* 1992; 70, 304-309.

13. Warde P, Mason M, Ding K, Kirkbride P, Brundage M, Cowan R, Gospodarowicz M, Sanders K, Kostachuk E, Swanson G, Barber J, Hiltz A, Parmar MKB, Sathya J, Anderson J, Hayter C, Hetherington J, Sydes MR, Parulekar W. Combined Androgen Deprivation Therapy and Radiation Therapy for Locally Advanced Prostate Cancer: A Randomised Phase 3 Trial. *The Lancet* 2011; DOI:10.1016/S0140-6736(11)61095-7.

14. Widmark A, Klepp O, Solberg A, Damber J, Angelsen A, Fransson P, Lund J, Tasdemir I, Hoyer M, Wiklund F, Fossa SD. Endocrine treatment, with or without radiotherapy, in locally advanced prostate cancer (SPCG-7/SFUO-3): an open randomized phase III trial. *Lancet* 2009; 373: 301-08

15. Schwartz C, Sprangers MAG. An Introduction to Quality of Life Assessment in Oncology: The Value of Measuring Patient Reported Outcomes. *Am J Manag Care* 2002; 8:S550-S559.

16. Bakas T, McLennon SM, Carpenter JS, Buelow JM, Otte JL, Hanna KM, Ellet ML, Hadler KA, Welch JL. Systematic Review of Health-Related Quality of Life Models. *Health and Quality of Life Outcomes* 2012, 10:134

17. Wilson IB, Cleary PD. Linking Clinical Variables with Health-Related Quality of Life. *JAMA* 1995; 273(1):59-65

18. Ferrans CE, Zerwic JJ, Wilbur JE, Larson JL. Conceptual Model of Health-Related Quality of Life. *J Nurs Scholarsh* 2005. 37(4):336-342

19. Guyatt GH, Feeny DH, Patrick DL. Measuring Health-related Quality of Life. *Annals of Internal Medicine* 1993; 118:622-629.

20. Alibhai SMH, Breunis H, Timilshina N, Johnston C, Tomlinson G, Tannock I, Krahn M, Fleshner NE, Warde P, Canning SD, Klotz L and Naglie G. Impact of Androgen - Deprivation Therapy on Physical Function and Quality of Life in Men With Nonmetastatic Prostate Cancer. *Journal of Clinical Oncology* 2010; 28,34: 5038-5045.

21. Wei JT, Dunn RL, Sandler HM, McLaughlin PW, Montie JE, Litwin MS, Nyquist L, and Sanda MG. Comprehensive Comparison of Health-Related Quality of Life After Contemporary Therapies for Localized Prostate Cancer. *Journal of Clinical Oncology* 2002; 20, 2: 557-566.

22. Eton DT and Lepore SJ. Prostate Cancer and Health-related Quality of Life: A Review of the Literature. *Psycho-Oncology* 2002; 11: 307–326.

23. Spilker, B. *Quality of Life Assessments in Clinical Trials*. New York: Raven Press, 1990.

24. Aday, LA and Cornelius, LJ. *Designing and Conducting Health Surveys: A Comprehensive Guide*. San Franciso: Jossey-Bass, 2006.

25. Aaronson NK, Beckmann JH. *The quality of life of the cancer patient*. New York: Raven Press, 1987.

26. Aaronson NK, Ahmedzai S, Berman B, et al. The Eurpean Organization for Research and Treatment of Cancer QLQ-C30: A quality-of-life instrument for use in international clinical trials in oncology. *J Natl Cancer Inst* 1993; 85: 365-376

27. Fayers P, Bottomley A. Quality of life research within the EORTC – the EORTC QLQ-C30. *European Journal of Cancer* 2002; 38, S125-S133.

28. Cella DF, Tulsky DS, Gray G, Sarafian B, Linn E, Bonomi A, Silberman M, Yellen SB, Winicour P, Brannon J, Eckberg K, Lloyd S, Purl S, Blendowski C, Goodman M, Barnicle M, Stewart I, McHale M, Bonomi P, Kaplan E, Taylor IV S, Thomas Jr, CR, and Harris J.

The Functional Assessment of Cancer Therapy Scale: Development and Validation of the General Measure. *J Clin Oncol* 1993; 11:570-579.

29. Esper P, Mo F, Chodak G, Sinner M, Cella D, Pienta KJ. Measuring Quality of Life in Men with Prostate Cancer using the Functional Assessment of Cancer Therapy-Prostate Instrument. *Urology* 1997; 50 (6): 920-928.

30. Uwer L, Rotonda C, Guillemin F, Miny J, Kaminsky MC, Mercier M, Tournier-Rangeard L, Leonard I, Montcuquet P, Rauch P and Conroy T. Responsiveness of EORTC QLQ-C30, QLQ-CR38 and FACT-C quality of life questionnaires in patients with colorectal cancer. *Health and Quality of Life Outcomes* 2011; 9:70.

31. Sprangers MAG, Moinpour CM, Moynihan TJ, Patrick DL, Revicki DA. Assessing Meaningful Change in Quality of Life Over Time: A Users' Guide for Clinicians. *Mayo Clin Proc* 2002; 77:561-571.

32. Lohr KN, Aaronson NK, Alonso J, Burnam MA, Patrick DL, Perrin EB, et al. Evaluating quality-of-life and health status instruments: development of scientific review criteria. *Clin Ther* 1996; 18:979-92

33. Osoba D, Rodrigues G, Myles J, Zee B and Pater J. Interpreting the Significance of Changes in Health-Related Quality of Life Scores. Journal of Clinical Oncology 1998; 16, 1: 139-144

34. Sloan J, Symonds T, Vargas-Chanes D, Fridley B. Practical Guidelines for Assessing the Clinical Significance of Health-Related Quality of Life Changes within Clinical Trials. *Drug Information Journal* 2003; 37, 1:23-31.

35. Osoba D, Bezjak A, Brundage M, Zee B, Tu D, Pater J. Analysis and Interpretation of health-related quality-of-life data from clinical trials: basic approach of the National Cancer Institute of Canada Clinical Trials Group. *Eur J Cancer* 2005; 41: 280-287

36. Cella D, Nichol M, Eton D, Nelson J, Mulani P. Estimating Clinically meaningful changes for the Functional Assessment of Cancer Therapy- Prostate: Results from a Clinical Trial of Patients with Metastatic Hormone-Refractory Prostate Cancer. *Value Health* 2009; 12:124-29

37. Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR, Clinical Significance Consensus Meeting G. Methods to explain the clinical significance of health status measures. *Mayo Clin Proc* 2002 Apr;77(4):371-383.

38. Cella D, Hahn EA, Dineen K. Meaningful change in cancer-specific quality of life scores: differences between improvement and worsening. *Qual Life Res* 2002, 11, 207–221.

39. King MT. The interpretation of scores from the EORTC quality of life questionnaire QLQ-C30. *Qual Life Res* 1996, 5, 555–567.

40. Norman GR, Sloan JA and Wyrwich KW. Interpretation of Changes in Health-related Quality of Life. The Remarkable Universality of Half a Standard Deviation. *Medical Care* 2003; 41, 5:582-592

41. Maringwa J, Quinten C, King M, Ringash J, Osoba D, Coens C, Martinelli F, Reeve BB, Gotay C, Greimel E, Flechtner H, Cleeland CS, Schmucker-Von Koch J, Weis J, Van Den Bent MJ, Stupp R, Taphoorn MJ, Bottomley A on behalf of the EORTC PROBE Project and

Brain Cancer Group. Minimal clinically meaningful differences for the EORTC QLQ-C30 and EORTC QLQ-BN20 scales in brain cancer patients. *Annals of Oncology* 2011; 22:2107-2112.

42. Ringash J, O'Sullivan B, Bezjak A, Redelmeier DA. Interpreting Clinically Significant Changes in Patient-Reported Outcomes. *Cancer* 2007; 110:196-202

43. Bedard G, Zeng L, Zhang L, Lauzon N, Holden L, Tsao M, Danjoux C, Barnes E, Sahgal A, Poon M, Chow E. Minimal important differences in the EORTC QLQ-C30 in patients with advanced cancer. *Asia-Pacific Journal of Clinical Oncology* 2014: 10:109-117.

44. Luckett T, King MT, Butow PN, Oguchi M, Rankin N, Price MA, Hackl NA and Heading G. Choosing between the EORTC QLQ-C30 and FACT-G for measuring health-related quality of life in cancer clinical research: issues, evidence and recommendations. *Annals of Oncology* 2011; 22: 2179-2190.

45. Sharp LK, Knight S, Nadler R et al. Quality of life in low-income patients with metastatic prostate cancer: divergent and convergent validity of three instruments. Qual Life Res 1999; 8(5): 461–470.

46. Blazeby JM, Kavadas V, Vickery CW et al. A prospective comparison of quality oflife measures for patients with esophageal cancer. Qual Life Res 2005; 14(2):387–393.

47. Holzner B, Kemmler G, Sperner-Unterweger B et al. Quality of life measurement in oncology—a matter of the assessment instrument? Eur J Cancer 2001; 37(18): 2349–2356.

48. Darling G, Eton DT, Sulman J et al. Validation of the functional assessment of cancer therapy esophageal cancer subscale. Cancer 2006; 107(4): 854–863.

49. Silpakit C, Sirilerttrakul S, Jirajarus M et al. The European Organization for Research and Treatment of Cancer Quality of Life Questionnaire (EORTC QLQC30): validation study of the Thai version. Qual Life Res. 2006; 15(1): 167–172.

50. Kemmler G, Holzner B, Kopp M et al. Comparison of two quality-of-life instruments for cancer patients: the functional assessment of cancer therapy general and the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire-C30. J Clin Oncol 1999; 17(9): 2932–2940.

51. Osoba D, Zee B, Pater J, Warr D, Kaizer L, Latreille J. Psychometric properties and responsiveness of the EORTC Quality of Life Questionnaire (QLQ-C30) in patients with breast, ovarian and lung cancer. *Quality of Life Research* 1994; 3: 353-364.

52. T. Conroy T, Mercier M, Bonneterre J, Luporsi E, Lefebvre JL, Lapeyre M, Puyraveau M, Schraub S. French version of FACT-G: Validation and comparison with other cancer-specific instruments. *European Journal of Cancer* 2004; 40: 2243-2252.

53. Katz JN, Phillips CB, Fossel AH, Liang MH: Stability and responsiveness of utility measures. Med Care 1994, 32:183-188.

54. King M, Bell ML, Costa D, Butow P, Oh B. The Quality of Life Questionnaire Core 30 (QLQ-C30) and Functional Assessment of Cancer-General (FACT-G) differ in

responsiveness, relative efficiency, and therefore required sample size. *Journal of Clinical Epidemiology* 2014; 67:100-107

55. Campbell MK, Elbourne DR, Altman DG for the CONSORT Group. CONSORT statement: extension to cluster randomized trials. BMJ 2004; 328:702-708

56. Hahn S, Puffer S, Torgerson D, Watson J. Methodological bias in cluster randomized trials. *BMC Medical Research Methodology* 2005; 5(1): 10

57. Calvert M, Blazeby J, Altman DG, Revicki DA, Moher D, Brundage MD. Reporting of Patient-Reported Outcomes in Randomized Trial - The CONSORT PRO Extension. JAMA 2014; 309(8): 814-822.

APPENDIX A

The EORTC QLQ-C30+3 and PR17 trial specific checklist

This box to be completed by the clinic nurse or clinical research associate: Pt. Serial #: _____ Pt. Initials: ___ ___ ___

European Organization for Research and Treatment of Cancer (EORTC)

Quality of Life Questionnaire (PR.3)

We are interested in some things about you and your health. Please answer all the questions yourself 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.

		No	Yes
1.	Do you have any trouble doing strenuous activities, like carrying a heavy shopping bag or a suitcase?	1	2
2.	Do you have any trouble taking a long walk?	1	2
3.	Do you have any trouble taking a short walk outside of the house?	1	2
4.	Do you have to stay in a bed or a chair for most of the day?	1	2
5.	Do you need help with eating, dressing, washing yourself or using the toilet?	1	2
6.	Are you limited in any way in doing either your work or doing household jobs?	1	2
7.	Are you completely unable to work at a job or to do household jobs?	1	2

During the past week:		Not at All	A Little	Quite a Bit	Very Much
8.	Were you short of breath?	1	2	3	4
9.	Have you had pain?	1	2	3	4
10.	Did you need to rest?	1	2	3	4
11.	Have you had trouble sleeping?	1	2	3	4
12.	Have you felt weak?	1	2	3	4
13.	Have you lacked appetite?	1	2	3	4
14.	Have you felt nauseated?	1	2	3	4

 (Core 30 Questionnaire © copyright 1992 EORTC Study Group on Quality of Life. All rights reserved.)

Please go on to the next page

This box to be completed by the clinic nurse or clinical research associate: Pt. Serial #: _____ Pt. Initials: ___ ___ ___

During the past week:	Not at All	A Little	Quite a Bit	Very Much
15. Have you vomited?	1	2	3	4
16. Have you been constipated?	1	2	3	4
17. Have you had diarrhea?	1	2	3	4
18. Were you tired?	1	2	3	4
19. Did pain interfere with your daily activities?	1	2	3	4
20. Have you had difficulty in concentrating on things like reading a newspaper or watching television?	1	2	3	4
21. Did you feel tense?	1	2	3	4
22. Did you worry?	1	2	3	4
23. Did you feel irritable?	1	2	3	4
24. Did you feel depressed?	1	2	3	4
25. Have you had difficulty remembering things?	1	2	3	4
26. Are you limited in doing either your work or household jobs?	1	2	3	4
27. Are you limited in pursuing your hobbies or other leisure time activities?	1	2	3	4
28. Has your physical condition or medical treatment interfered with your family life?	1	2	3	4
29. Has your physical condition or medical treatment interfered with your social activities?	1	2	3	4
30. Has your physical condition or medical treatment caused you financial difficulties?	1	2	3	4

This box to be completed by the clinic nurse or clinical research associate: Pt. Serial #: _____ Pt. Initials: ___ ___ ___

For the following questions please circle the number between 1 and 7 that best applies to you.

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

1	2	3	4	5	6	7
Very Poor						Excellent

32. How would you rate your overall health during the past week?

1	2	3	4	5	6	7
Very Poor						Excellent

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

1	2	3	4	5	6	7
Very Poor						Excellent

Patients sometimes report that they have the following symptoms. Please indicate the extent to which you have experienced these symptoms during the past week, by circling the number that best applies to you.

During the past week:	**Not at All**	**A Little**	**Quite a Bit**	**Very Much**
34. Did you have to pass urine more frequently than normal for you?	1	2	3	4
35. Did you have difficulty passing your urine?	1	2	3	4
36. Did you have pain when you passed urine?	1	2	3	4
37. Did you have blood in your urine?	1	2	3	4
38. Did you have difficulty emptying your bladder completely?	1	2	3	4
39. Did you have difficulty controlling your urination (for example dribbling)?	1	2	3	4
40. Did you have accidental wetting of your underwear?	1	2	3	4
41. Did you have to wear added protection to prevent accidental wetting of your underwear?	1	2	3	4

This box to be completed by the clinic nurse or clinical research associate: Pt. Serial #: _____ Pt. Initials: ___ ___ ___

During the past week:	Not at All	A Little	Quite a Bit	Very Much
42. Did you have any bleeding from your rectum (for example, with a bowel movement)?	1	2	3	4
43. Did you have any pain in your rectum?	1	2	3	4
44. Did you have hot flashes?	1	2	3	4
45. Did you have any bothersome breast enlargement?	1	2	3	4
46. Has your present condition affected your sex life?	1	2	3	4
47. Did you limit your activities outside your home?	1	2	3	4
If you answered "not at all", please proceed to question 49.				
48. If you limited your activities outside the home, was this because of your urination problems?	1	2	3	4
49. Did you have to get up at night to pass urine?	1	2	3	4
If you answered "not at all", please proceed to next page.				
50. How much did getting up at night interfere with your sleep?	1	2	3	4

APPENDIX B

The FACT-P

This box to be completed by the clinic nurse or clinical research associate: Pt. Serial #: _____ Pt. Initials: ___ ___ ___

Eastern Cooperative Oncology Group (ECOG) FACT-P (version 2)

Quality of Life Questionnaire (PR.3)

Below is a list of statements that other people with your illness have said are important. By circling one number per line, please indicate how true each statement has been for you during the past 7 days.

During the past 7 days:	Circle one number				
Physical Well-being	**not at all**	**a little bit**	**some-what**	**quite a bit**	**very much**
1. I have a lack of energy	0	1	2	3	4
2. I have nausea	0	1	2	3	4
3. I have trouble meeting the needs of my family . .	0	1	2	3	4
4. I have pain	0	1	2	3	4
5. I am bothered by side effects of treatment	0	1	2	3	4
6. In general, I feel sick	0	1	2	3	4
7. I am forced to spend time in bed	0	1	2	3	4

8. How much does your physical well-being affect your quality of life?

Not at all 0 1 2 3 4 5 6 7 8 9 10 Very much so

This box to be completed by the clinic nurse or clinical research associate: Pt. Serial #: _____ Pt. Initials: ___ ___ ___

During the past 7 days: Circle one number

Social/Family Well-Being	not at all	a little bit	some-what	quite a bit	very much
9. I feel distant from my friends	0	1	2	3	4
10. I get emotional support from my family	0	1	2	3	4
11. I get support from my friends and neighbours . . .	0	1	2	3	4
12. My family has accepted my illness	0	1	2	3	4
13. Family communication about my illness is poor . .	0	1	2	3	4

If you have a spouse/partner, or are sexually active, please answer # 14-15. Otherwise go to # 16.

14. I feel close to my partner (or main support)	0	1	2	3	4
15. I am satisfied with my sex life	0	1	2	3	4

16. How much does your social/family well-being affect your quality of life?

Not at all 0 1 2 3 4 5 6 7 8 9 10 Very much so

During the past 7 days: Circle one number

Relationship With Doctor	not at all	a little bit	some-what	quite a bit	very much
17. I have confidence in my doctor(s)	0	1	2	3	4
18. My doctor is available to answer my questions . .	0	1	2	3	4

19. How much does your relationship with the doctor affect your quality of life?

Not at all 0 1 2 3 4 5 6 7 8 9 10 Very much so

This box to be completed by the clinic nurse or clinical research associate: Pt. Serial #: _____ Pt. Initials: ___ ___ ___

During the past 7 days: Circle one number

Emotional Well-Being	not at all	a little bit	some-what	quite a bit	very much
20. I feel sad	0	1	2	3	4
21. I am proud of how I'm coping with my illness	0	1	2	3	4
22. I am losing hope in the fight against my illness	0	1	2	3	4
23. I feel nervous	0	1	2	3	4
24. I worry about dying	0	1	2	3	4

25. How much does your emotional well-being affect your quality of life?

Not at all 0 1 2 3 4 5 6 7 8 9 10 Very much so

During the past 7 days: Circle one number

Functional Well-Being	not at all	a little bit	some-what	quite a bit	very much
26. I am able to work (include work in home)	0	1	2	3	4
27. My work (including work in home) is fulfilling	0	1	2	3	4
28. I am able to enjoy life "in the moment"	0	1	2	3	4
29. I have accepted my illness	0	1	2	3	4
30. I am sleeping well	0	1	2	3	4
31. I am enjoying my usual leisure pursuits	0	1	2	3	4
32. I am content with the quality of my life right now	0	1	2	3	4

33. How much does your functional well-being affect your quality of life?

Not at all 0 1 2 3 4 5 6 7 8 9 10 Very much so

This box to be completed by the clinic nurse or clinical research associate: Pt. Serial #: _____ Pt. Initials: ___ ___ ___

During the past 7 days: Circle one number

Additional Concerns	not at all	a little bit	some-what	quite a bit	very much
34. I am losing weight	0	1	2	3	4
35. I have a good appetite	0	1	2	3	4
36. I have aches and pains that bother me	0	1	2	3	4
37. I have certain areas of my body where I experience significant pain	0	1	2	3	4
38. My pain keeps me from doing things I want to do	0	1	2	3	4
39. I am satisfied with my present comfort level	0	1	2	3	4
40. I am able to feel like a man	0	1	2	3	4
41. I have trouble moving my bowels	0	1	2	3	4
42. I have difficulty urinating	0	1	2	3	4
43. I urinate more frequently than usual	0	1	2	3	4
44. My problems with urinating limit my activities	0	1	2	3	4
45. I am able to have and keep an erection	0	1	2	3	4

46. How much do these additional concerns affect your quality of life?

Not at all 0 1 2 3 4 5 6 7 8 9 10 Very much so

Please fill in your initials to indicate that you have completed this questionnaire: _____

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

Thank you.

APPENDIX C

Study Power

Objective 3B – Chi-square test

Estimating the minimal detectable effect with a given sample size and power level (*a priori* calculation)

n	α	power	p0	p1	Δp1-p0
311	0.05	80%	0.5	0.657	.157

Determining sample size needed to detect differences observed (post-hoc calculations)

Functional domain	n	α	power	p0	p1	Δp1-p0
Physical domain	1582	0.05	80%	0.52	0.59	0.07
Role/Functional domain	1208	0.05	80%	0.52	0.6	0.08
Emotional domain	4038	0.05	80%	0.27	0.31	0.04
Social domain	77538	0.05	80%	0.55	0.56	0.01
Fatigue	3866	0.05	80%	0.75	0.71	0.04

Determining the power for observed differences in proportions (post-hoc calculations)

Functional domain	n	α	power	p0	p1	Δp1-p0
Physical domain	311	0.05	24%	0.52	0.59	0.07
Role/Functional domain	311	0.05	29%	0.52	0.6	0.08
Emotional domain	311	0.05	12%	0.27	0.31	0.04
Social domain	311	0.05	4%	0.55	0.56	0.01
Fatigue	311	0.05	12%	0.75	0.71	0.04

Where n is sample size
Where α is type 1 error
Where power is the probability of correctly rejecting the null hypothesis
Where p0 is the estimated proportion of patients with worsened HRQL in one of instrument groups
Where p1 is the detectable proportion of patients with worsened HRQL in the other instrument group
Where Δp1-p0 is the detectable difference in proportions

Objective 3C – Cox Proportional Hazards Regression

Determining sample size needed to detect hazard ratios observed (post-hoc calculations)

Functional domain	n	α	power	HR
Physical domain	144	0.05	80%	0.623
Role/Functional domain	214	0.05	80%	0.679
Emotional domain	616	0.05	80%	0.796
Social domain	658	0.05	80%	0.802
Fatigue	7817	0.05	80%	0.938

Determining the power for observed hazard ratios (post-hoc calculations)

Functional domain	n	α	power	HR
Physical domain	311	0.05	99%	0.623
Role/Functional domain	311	0.05	92%	0.679
Emotional domain	311	0.05	51%	0.796
Social domain	311	0.05	49%	0.802
Fatigue	311	0.05	8%	0.938

Where n is sample size
Where α is type 1 error
Where power is the probability of correctly rejecting the null hypothesis
Where HR is the hazard ratio

APPENDIX D

Construction of Domain/Item Scores for the EORTC QLQ C30+3/ PR.17 vs. the FACT-P

	EORTC QLQ-C30+3/PR.17			FACT-P		
	Domain/Item	**No. of items**	**No. of response categories for each item**	**Domain/Item**	**No. of items**	**No. of response categories for each item**
Functional scales	Physical functioning	7	2	Physical Well being	7	5
	Role functioning	2	4	Functional Well being	7	5
	Emotional functioning	4	4	Emotional Well being	5	5
	Social functioning	2	4	Social/Family Well being	7	5
Symptom scales/items	Fatigue	3	4	Fatigue	1	5
	Constipation	1	4			
	Diarrhea	1	4			
	Urine symptoms (PR.17)	8	4	Urine problems	3	5
	Bowel Rectum symptoms (PR.17)	2	4	Bowel trouble	1	5
	Urination at night (PR.17)	1	4			

APPENDIX E

Symptom Domain Construction for the EORTC QLQ-PR25 vs the PR17 Trial Specific Checklist

	PR25	PR17
Urine Symptom Domain	31. Have you had to urinate frequently during the day?	34. Did you have to pass urine more frequently than normal for you?
	32. Have you had to urinate frequently at night?	35. Did you have difficulty passing your urine?
	33. When you felt the urge to pass urine, did you have to hurry to get to the toilet?	36. Did you have pain when you passed urine?
	34. Was it difficult for you to get enough sleep, because you needed to get up frequently at night to urinate?	37. Did you have blood in your urine?
	35. Have you had difficulty going out of the house because you needed to be close to a toilet?	38. Did you have difficulty emptying your bladder completely?
	36. Have you had any unintentional release (leakage) of urine?	39. Did you have difficulty controlling your urination (for example dribbling)?
	37. Did you have pain when you urinated?	40. Did you have accidental wetting of your underwear?
	38. Answer this question only if you wear an incontinence aid. Has wearing an incontinence aid been a problem for you?	41. Did you have to wear added protection to prevent accidental wetting of your underwear?
	39. Have your daily activities been limited by your urinary problems?	
Bowel/Rectum symptom domain	40. Have your daily activities been limited by your bowel problems?	42. Did you have any bleeding from your rectum (for example, with a bowel movement)?
	41. Have you had any unintentional release (leakage) of stools?	43. Did you have any pain in your rectum?
	42. Have you had blood in your stools?	
	43. Did you have a bloated feeling in your abdomen?	

APPENDIX F

Ethics Approval

QUEEN'S UNIVERSITY HEALTH SCIENCES & AFFILIATED TEACHING HOSPITALS RESEARCH ETHICS BOARD-DELEGATED REVIEW
November 27, 2013

Ms. Yvonne Murray
Department of Public Health Sciences
Queen's Cancer Research Institute

Dear Ms. Murray
Study Title: EPID-451-13 Comparison of Two Quality of Life Instruments in a Phase III Randomized Clinical Trial of Men with Prostate Cancer (NCIC CTG PR.3): The QLQ-C30+3 with Trial Specific Checklist (PR.17) versus the FACT-P
File # 6011423
Co-Investigators: Dr. M. Brundage, Dr. B. Chen, Dr. H. Richardson

I am writing to acknowledge receipt of your recent ethics submission. We have examined the protocol for your project (as stated above) and consider it to be ethically acceptable. This approval is valid for one year from the date of the Chair's signature below. This approval will be reported to the Research Ethics Board. Please attend carefully to the following listing of ethics requirements you must fulfill over the course of your study:

Reporting of Amendments: If there are any changes to your study (e.g. consent, protocol, study procedures, etc.), you must submit an amendment to the Research Ethics Board for approval. Please use event form: HSREB Multi-Use Amendment/Full Board Renewal Form associated with your post review file # **6011423** in your Researcher Portal (https://eservices.queensu.ca/romeo_researcher/)

Reporting of Serious Adverse Events: Any unexpected serious adverse event occurring locally must be reported within 2 working days or earlier if required by the study sponsor. All other serious adverse events must be reported within 15 days after becoming aware of the information. Serious Adverse Event forms are located with your post-review file **6011423** in your Researcher Portal (https://eservices.queensu.ca/romeo_researcher/)

Reporting of Complaints: Any complaints made by participants or persons acting on behalf of participants must be reported to the Research Ethics Board within 7 days of becoming aware of the complaint. Note: All documents supplied to participants must have the contact information for the Research Ethics Board.

Annual Renewal: Prior to the expiration of your approval (which is one year from the date of the Chair's signature below), you will be reminded to submit your renewal form along with any new changes or amendments you wish to make to your study. If there have been no major changes to your protocol, your approval may be renewed for another year.

Yours sincerely,

Albert J. Clark.

Chair, Health Sciences Research Ethics Board
November 27, 2013

Investigators please note that if your trial is registered by the sponsor, you must take responsibility to ensure that the registration information is accurate and complete

QUEEN'S UNIVERSITY HEALTH SCIENCES & AFFILIATED TEACHING HOSPITALS RESEARCH ETHICS BOARD

The membership of this Research Ethics Board complies with the membership requirements for Research Ethics Boards and operates in compliance with the Tri-Council Policy Statement; Part C Division 5 of the Food and Drug Regulations, OHRP, and U.S DHHS Code of Federal Regulations Title 45, Part 46 and carries out its functions in a manner consistent with Good Clinical Practices.

Federalwide Assurance Number: #FWA00004184, #IRB00001173

Current 2013 membership of the Queen's University Health Sciences & Affiliated Teaching Hospitals Research Ethics Board:

Dr. A.F. Clark, Emeritus Professor, Department of Biomedical and Molecular Sciences, Queen's University (Chair)

Dr. H. Abdollah, Professor, Department of Medicine, Queen's University

Dr. R. Brison, Professor, Department of Emergency Medicine, Queen's University

Dr. C. Cline, Assistant Professor, Department of Medicine, Director, Office of Bioethics, Queen's University, Clinical Ethicist, Kingston General Hospital

Dr. M. Evans, Community Member

Ms. J. Hudacin, Community Member

Dr. B. Kisilevsky, Professor, School of Nursing, Departments of Psychology and Obstetrics and Gynaecology, Queens's University

Mr. D. McNaughton, Community Member

Ms. P. Newman, Pharmacist, Clinical Care Specialist and Clinical Lead, Quality and Safety, Pharmacy Services, Kingston General Hospital

Ms. S. Rohland, Privacy Officer, ICES-Queen's Health Services Research Facility, Research Associate, Division of Cancer Care and Epidemiology, Queen's Cancer Research Institute

Dr. A. Singh, Professor, Department of Psychiatry, Queen's University

Dr. J. Walia, Assistant Professor and Clinical Geneticist, Department of Paediatrics, Queen's University and Kingston General Hospital

Ms. K. Weisbaum, LL.B. and Adjunct Instructor, Department of Family Medicine (Bioethics)