Improving decision-making:
Deriving patient-valued utilities from a disease-specific quality of life questionnaire for evaluating clinical trials

by

Peter S Grimison

This thesis is submitted in full satisfaction of the requirements for the degree of Doctor of Philosophy, University of Sydney, September 2009

NHMRC Clinical Trials Centre,
School of Public Health,
Faculty of Medicine,
University of Sydney

Abstract

The aim of the work reported in this thesis was to develop a scoring algorithm that converts ratings from a validated disease-specific quality of life questionnaire called the Utility-Based Questionnaire-Cancer (UBQ-C) into a utility index that is designed for evaluating clinical trials to inform clinical decisions about cancer treatments.

The UBQ-C includes a scale for global health status (1 item); and subscales for physical function (3 items), social/usual activities (4 items), self-care (1 item), and distresses due to physical and psychological symptoms (21 items). Data from three studies was used. A valuation survey consisted of patients with advanced cancer (n=204) who completed the UBQ-C and assigned time-trade-off utilities about their own health state. Clinical trials were of chemotherapy for advanced (n=325) and early (n=126) breast cancer. A scoring algorithm was derived to convert the subscales into a subset index, and combine it with the global scale into an overall quality of life index, which was converted to a utility index with a power transformation. Optimal weights were assigned to the subscales that reflected their correlations with a global scale in each study. The derived utilities were validated by comparison with other patient characteristics. Each trial was evaluated in terms of differences in utility between treatment groups.

In the valuation survey, the weights (range 0 to 1) for the subset index were: physical function 0.28, social/usual activities 0.06, self-care 0.01, and distresses 0.64. Weights for the overall quality of life index were health status 0.65 and subset index 0.35. The mean of the utility index scores was similar to the mean of the time trade-off utilities (0.92 vs. 0.91, $p=0.6$). The weights were adjusted in each clinical trial. The utility index was substantially correlated with other measures of quality of life, discriminated between breast cancer that was advanced rather than early (means 0.88 vs 0.94, $p<0.0001$), and was responsive to toxic effects of chemotherapy in early breast cancer (mean change 0.07, $p<0.0001$). There were trends to better mean scores on the utility index for patients allocated to standard-dose versus high-dose chemotherapy in the early cancer trial ($p=0.1$), and oral versus intravenous chemotherapy in the advanced cancer trial ($p=0.2$).

In conclusion, data from a simple, self-rated, disease-specific questionnaire can be converted into a utility index based on cancer patients' preferences. The index can be optimised in different clinical contexts to reflect the relative importance of different aspects of quality of life to the patients in a trial. The index can be used to generate utility scores and quality-adjusted life-years in clinical trials. It enables the evaluation of the net effect of treatments on health-related quality of life (accounting for trade-offs between disparate aspects); the evaluation of the net benefit of treatments (accounting for trade-offs between quality of life and survival); and an alternate perspective for comparing the incremental cost-effectiveness of treatments (accounting for trade-offs between net benefit and costs).

The practical significance of this work is to facilitate the integration of data about health-related quality of life with traditional trial endpoints such as survival and tumour response. This will better inform clinical decision-making, and provide an alternate viewpoint for economic decision-making. Broadly, it will help patients, clinicians and health funders make better decisions about cancer treatments, by considering potential trade-offs between effects on survival and health-related quality of life.

Dedicated to

Lindsay Grimison

1947 to 2007

Preface

I undertook this thesis as a full-time PhD student at the NHMRC Clinical Trials Centre within the School of Public Health of the University of Sydney, under the supervision of A/Prof Martin R Stockler and Prof R John Simes.

Author's Contribution

I, Peter Grimison, was primarily and principally responsible for the following: development of the research proposal; submission of scholarship applications to support the research; selection of research methods; data analysis; interpretation of the findings; and drafting the thesis.

My supervisors were responsible for the concept of the project, provided constructive feedback and critique through out all stages of the project, and reviewed multiple drafts of this thesis. Vicki Greatorex and Andrew Martin did data collection for the valuation survey described in chapter 3 and analysed in chapter 5. The investigators and research staff of the Australia New Zealand Breast Cancer Trials Group and the International Breast Cancer Study Group were responsible for trial design, data collection and data management of the clinical trials described in chapter 3 and analysed in chapters 6 and 7. Prof Malcolm Hudson contributed to selection of statistical techniques and interpretation of statistical analyses for the work described in chapters 4 to 7, and reviewed drafts of chapters 5 and 6.

Ethical Clearances

The valuation survey, that was presented in chapter 3 and analysed in chapter 5, was approved by the Ethics Review Committee of Royal Prince Alfred Hospital, Sydney (Protocol N^{o} 940069) and the Human Research Ethics Committee at the University of Sydney (Ref N^{o} 11-2006/9626). The multi-centre clinical trials for early and advanced breast cancer, that were presented in chapter 3 and analysed in chapters 6 to 7, were approved by the Human Research Ethics Committees at all participating institutions.

Publications arising

The following published peer-reviewed journal article and conference abstracts are a direct result of the research undertaken in this thesis:

Peer-reviewed Journal Articles

1. **Grimison PS**, Simes RJ, Hudson HM, Stockler MR.
   Deriving a patient-based utility index from a cancer-specific quality of life questionnaire. Value in Health 2009; 12(5):800-807 (PMID 19508665).

2. **Grimison PS**, Simes RJ, Hudson HM, Stockler MR.
   Preliminary validation of an optimally-weighted patient-based utility index by application to randomised trials in breast cancer. Value in Health 2009; 12(6): 967-976 (PMID 10490566).

Published Abstracts

1. **Grimison P**, Simes J, Stockler M. Deriving valid utilities for comparing treatments in clinical trials using standard quality of life questionnaires.
   (i) Asia-Pacific Journal of Clinical Oncology (Clinical Oncological Society of Australia 2006 Annual Scientific Meeting Proceedings) 2007; 3(1):55
   Awarded best oral presentation
   (ii) Journal of Clinical Oncology (American Society of Clinical Oncology Annual Meeting Proceedings) 2007; 25(18S):6500

2. **Grimison P.S**., Simes, R.J., Stockler M.R. Establishing the validity and precision of a weighted global measure of health-related quality of life (HRQL) for a cancer-specific questionnaire using data from a randomised trial for advanced breast cancer.
   Asia-Pacific Journal of Clinical Oncology (Medical Oncology Group of Australia Annual Scientific Meeting Proceedings) 2007; 3(1S):A22

3. **Grimison P.S**., Simes, R.J., Stockler M.R. Development and validation of optimally weighted measures of global health-related quality of life and utility based on a cancer-specific quality of life instrument. Value in Health (International Society for Pharmacoeconomics and Outcomes Research 10th Annual European Congress Proceedings) 2007; 10(6):A226
   Awarded best student podium presentation

Ongoing work arising from the research undertaken in this thesis

The following conference abstracts report the analyses of data from randomised controlled trials using the methods developed in this thesis. Peer-reviewed journal articles are planned for each study:

Published Abstracts

1. Stockler M. R., **Grimison P. S**., Price T. J., van Hazel G. A., Robinson B. A., Broad A., Ganju V., Wilson K., Tunney V., Tebbutt N. C., Australasian Gastro-Intestinal Trials Group. Comparing utilities for advanced colorectal cancer valued from societal and cancer-patients' perspectives using baseline data from the MAX study
   (i) Journal of Clinical Oncology (ASCO Annual Meeting Proceedings) 2008; 25(15S):6504
   (ii) Asia-Pacific Journal of Clinical Oncology (COSA ASM Proceedings) 2007; (S2):A104.
   (iii) Asia-Pacific Journal of Clinical Oncology (MOGA ASM Proceedings) 2008; 4(S1):A16

2. **Grimison P. S.**, Coates A. S., Forbes J. F., Cuzick J., Furnival C., Craft P. S., Snyder R. D., Thornton R., Lindsay D. F., Simes R. J., Australian New Zealand Breast Cancer Trials s Group. Tamoxifen (TAM) for the prevention of breast cancer: importance of specific aspects of health-related quality of life to global health status in the ANZ BCTG substudy of IBIS-1 (ANZ 92P1).
   (i) Journal of Clinical Oncology (ASCO Annual Meeting Proceedings) 2008; 25(15S):1516
   (ii) Asia-Pacific Journal of Clinical Oncology (MOGA ASM Proceedings) 2008; 4(S1):A9
   (iii) Asia-Pacific Journal of Clinical Oncology (COSA ASM Proceedings) 2008; 4(S2):A87

Current analyses are evaluating the primary endpoint of the following study, using results generated from chapter 7 of this thesis:

Published Abstract

1. Stockler M, Sourjina T, **Grimison P**, Gebski V, Byrne M, Harvey V, Francis P, Nowak A, Coates A, Forbes J.
   A randomized trial of capecitabine (C) given intermittently (IC) versus continuously (CC) versus classical CMF as first line chemotherapy for advanced breast cancer (ABC).
   Journal of Clinical Oncology (ASCO Annual Meeting Proceedings) 2007; 25(18S):1031

Acknowledgements

I would like to gratefully acknowledge the help and support of my Supervisors. I thank A/Prof Martin Stockler for his selflessness, encouragement, time, knowledge and broad expertise; from generating ideas through to the final draft of the thesis. I look forward to a long continued association. I thank Prof John Simes, who was responsible for much of the prior research at the NHMRC Clinical Trials Centre that enabled this research project to take place, conceived the project, and has provided careful advice based on extensive knowledge throughout my time here.

My colleagues at the NHMRC Clinical Trials Centre have also provided invaluable help and support. A particular thank you to Prof Malcolm Hudson for his statistical expertise and to Doctors Michaella Smith and Corona Gainford for their interest and encouragement. Thanks also to Prof Michael Friedlander who facilitated my contact with the NHMRC Clinical Trials Centre.

I am indebted to the patients that participated in the research. Many thanks also to Vicki Greatorex and Andrew Martin who obtained invaluable data by interviewing over 200 patients with advanced cancer, and to the investigators and research staff of the Australia New Zealand Breast Cancer Trials Group and the International Breast Cancer Study Group who recruited patients and generously provided trial data for analyses. Prof Alan Coates, Prof John Forbes and Dianne Lindsay from the Australia New Zealand Breast Cancer Trials Group deserve special mention.

My research was possible only through the support of a Medical Postgraduate Research Scholarship from the National Health and Medical Research Council, a Research Scholar Award from the Cancer Institute NSW, a Post Graduate Support Grant from GlaxoSmithKline Australia, and a Postgraduate Scholarship in Cancer Clinical Trials and Quality of Life from the NHMRC Clinical Trials Centre at the University of Sydney. For these I am most grateful.

A heartfelt thank you to Walter McIntosh, my family, and my friends who have provided their constant support over the years it has taken to produce this thesis.

Table of contents

List of Appendices

List of Tables

List of Figures

List of Abbreviations

HRQL	health-related quality of life
QALY	quality-adjusted life-year
UBQ-C	Utility-Based Questionnaire-Cancer

1. Introduction

1.1 Rationale and origins of the thesis

The choice of treatments for cancer is growing rapidly. Cancer treatments may extend life, relieve cancer symptoms or improve physical and psychological function. On the other hand, cancer treatments can also cause significant toxicity, inconvenience and other costs. Patients, clinicians and health funders need to know if the benefits are sufficient to outweigh the harms.

An index of net clinical benefit that explicitly weighs up these trade-offs is helpful for making these decisions. Such an index can be used to evaluate and compare treatments on a common scale that incorporates disparate treatment effects like gains in survival duration, improvements in health status and health-related quality of life (HRQL) due to relief of cancer symptoms on the one hand, and deteriorations in HRQL due to treatment-related side effects on the other.

The quality-adjusted life year (QALY) is one such index of net clinical benefit. The QALY approach combines effects on survival duration, expressed in life years, with net effects on HRQL, expressed as a 'utility' (defined in the next paragraph). This approach enables cancer treatments to be compared on a common scale. Analyses of cancer treatments in terms of utilities and QALYs are commonly used to inform economic decisions by funders and policy-makers, but can also be used to inform clinical decisions by patients and clinicians.

A utility is a single number expressing the net impact of a health condition and its treatment on HRQL. It represents a unified assessment about the desirability of a health state relative to full health (one) and death (zero). A utility can be directly elicited from a respondent using a standard gamble or time trade-off interview. This task is complex and resource-intensive. An alternative approach is to derive a utility index from an individual's responses to a simple self-rated questionnaire.

The work presented in this thesis was motivated by the desire to evaluate a series of randomised clinical trials conducted by the NHMRC Clinical Trials Centre at the University of Sydney, Australia using the QALY approach. For each trial, data about

the effects of treatments on survival duration and other time-to-event outcomes were obtained. Data about HRQL were obtained using a cancer-specific HRQL questionnaire called the Utility-Based Questionnaire-Cancer. This questionnaire was developed and validated at the NHMRC Clinical Trials Centre, and provided descriptive information about effects of treatments across a range of aspects of HRQL. The purpose of the work reported in this thesis was to further develop the questionnaire as a feasible method of obtaining valid and reliable utility scores. The utility scores could be used to inform analyses of the clinical trials using the QALY approach. This could aid in interpretation of clinical trial results, and ultimately could help to inform clinical decisions by patients and clinicians about cancer treatments..

1.2 Aim

The aim of the work reported in this thesis was to develop a scoring algorithm that converts the responses to a cancer-specific questionnaire into an optimally weighted utility index. The index is intended to:

i) reflect the perspective of patients with cancer;
ii) be optimally weighted for comparisons in specific clinical contexts, *and*
iii) be feasible for use in cancer clinical trials.

The utility index can be used to describe the net effect of cancer treatments on quality of life, and to evaluate trade-offs between quality and quantity of life using quality-adjusted survival analyses.

1.3 Main approaches

Source of data

Ambulatory patients with advanced cancer (n=204) assigned utilities for their current state of health in a face-to-face interview, and completed the Utility-Based Questionnaire-Cancer.

Participants in two randomised controlled trials of chemotherapy for breast cancer (n=421) also completed the Utility-Based Questionnaire-Cancer at baseline and during treatment. The first trial compared oral versus intravenous chemotherapy for advanced breast cancer. The second trial compared high-dose chemotherapy with stem cell support to standard adjuvant chemotherapy for high risk early breast cancer.

Algorithm development

A scoring algorithm was derived that converted the 30 items of the Utility-Based Questionnaire-Cancer into an optimally weighted index of overall HRQL using data from the cross-sectional study. The approach incorporated the views and preferences of trial patients for rating changes in aspects of health-related quality of life and weighing their importance.

A second equation was derived that converted the index of overall health-related quality of life to a utility index using data from the cross-sectional study. The best transformation was selected in terms of its predictive ability.

Validation and application

The utility index was validated using data from the two randomised controlled trials.

The system was extended so that it could be applied to longitudinal data, and then applied to evaluate the net benefit of treatments in terms of overall health-related quality of life and utility in the randomised controlled trial of chemotherapy for advanced breast cancer. The differences in overall HRQL and utility between treatment groups were determined, and utility weights were calculated to be integrated with survival data for quality-adjusted survival analyses.

1.4 Outline of chapters

Chapter 2 provides a general introduction to the research area that is built upon in later chapters. The first section introduces the concept of HRQL as an outcome in health care, and describes the relevance of quality of life data for informing both clinical decisions and economic decisions about cancer treatments. The middle sections describe the different approaches for measuring health-related quality of life: standard value-based health-related quality of life instruments, direct utility-based scaling methods, and utility-based instruments that incorporate elements of both. The third section reviews methods that have been used to construct scoring algorithms for utility-based instruments.

Chapter 3 describes the study materials used for the work presented in later chapters. The content and development of the Utility-Based Questionnaire-Cancer, and other questionnaires used for validation, are described. The study designs of the component studies used to develop, validate and apply the algorithm are also presented.

Chapter 4 gives an overview of all the methods used, and the background relating to the methodological approach taken. By necessity, some information in chapters 3 and 4 is repeated in the studies reported in chapters 5, 6 and 7.

Chapter 5 describes how a scoring algorithm was derived to convert the responses on the Utility-Based Questionnaire-Cancer into a utility index.

Chapter 6 describes the optimisation of the scoring algorithm in the setting of clinical trials for early breast cancer and advanced breast cancer, validation of the resultant utility index by comparison with related measures, and its application to evaluate the net effects of treatments tested in a randomised controlled trial of adjuvant chemotherapy for early breast cancer.

Chapter 7 describes the further development of the scoring algorithm for use with longitudinal data, and its application to evaluate the net effects of treatments tested in a randomised controlled trial of palliative chemotherapy for advanced breast cancer.

Chapter 8 revisits the rationale for, aims of, and approach taken to the thesis, and summarises the principal findings. The strengths and limitations of the work undertaken are discussed, the implications of the findings for clinical practice and future research are considered, and priorities for future research are identified. Finally, the contribution to knowledge and practical significance of the work are stated.

2. Background

2.1 Overview

This chapter provides a descriptive overview of methods used to assess health-related quality of life. Measures that focus on specific aspects of quality of life are compared to measures that focus on global quality of life, in terms of their differing measurement properties and applications. The chapter is focussed on methods to produce summary data from quality of life measures that can be applied to evaluate treatments in clinical trials, and more broadly to inform decision-making about treatments.

2.2 Health-related quality of life

2.2.1 Definition

Quality of life is a subjective and abstract concept that reflects an individual's perception of and response to their unique circumstances. It is a broad and multi-dimensional concept which can be separated into 'health-related' and 'non-health' related aspects. Health-related quality of life (HRQL) focuses on the potential effects of a disease and its treatment on quality of life. Non-health aspects of quality of life include the quality of the environment, an individual's standard of living, and political freedom. Health researchers tend to focus on HRQL, because non-health aspects are unlikely to be affected by disease or treatment [1-3].

No universally accepted definition of HRQL exists, but there is broad agreement that an assessment of HRQL should include physical, psychological and emotional, and social dimensions. Many researchers recommend that other dimensions be included, for example symptoms of disease and side effects of treatment, such as pain, nausea, sleep disturbance, cognitive impairment, or sexual dysfunction; and broader aspects of quality of life such as spirituality or patient satisfaction [2-7].

In this thesis, the symptoms of disease and side effects of treatment are included as dimensions of HRQL. This is commonly done by researchers, but there are theoretical problems with this approach. Physical, psychological, emotional and social dimensions can be considered as indicator items that indicate (or reflect) the effects of HRQL impairment [8]. Symptoms and side effects can be considered as causal items that cause impairment in a patient's HRQL. Measuring both indicator and causal items can lead to 'double-counting' of effects of HRQL. However it is intuitively felt that their inclusion is important in a questionnaire used to measure treatment effects in clinical trials. This approach is used by many researchers [2].

2.2.2 Relevance of HRQL to the evaluation of cancer treatments

Assessing HRQL is increasingly recognised as an important component of the evaluation of treatments for cancer and other conditions. HRQL information is used to guide individual patient care, to evaluate cancer treatments tested in groups of patients in clinical trials, and to inform decisions for populations by health funders

and policy makers. At the individual patient level, improvements (or deteriorations) in HRQL can be used to determine the net benefit of a treatment in a particular patient. At a clinical trial level, HRQL assessment helps to evaluate and compare average effects of treatments on specific aspects of HRQL, and the trade-offs of beneficial and harmful effects on different aspects of HRQL. HRQL assessment can also evaluate and compare treatments in terms of the trade-offs between quantity and quality of life by integrating HRQL data with survival data into a single common metric called quality-adjusted life years (QALYs). At a population level, health economists use QALY data about the effectiveness of treatments together with data about the costs of treatments in cost-effectiveness analysis to inform funding and policy decisions [2, 5, 9-10].

2.2.3 Overview of approaches for measuring HRQL

A diverse range of approaches for measuring HRQL exist, because no single approach is suitable for all situations. This reflects the broad and complex nature of the concept of HRQL, the variety of backgrounds of analysts, and the multiple purposes for HRQL assessment described above [11]. This section addresses the distinction between measures that are 'global' or 'specific', and measures that are derived by a 'value-based' or 'utility-based' scaling method.

One way of classifying HRQL measures is by whether the content of its questions are 'global' or 'specific'. A global question asks respondents for a unified assessment of HRQL or health status. Examples include rating scales such as the health status thermometer of the Utility-Based Questionnaire-Cancer [12-13] (appendix 1), and the Spitzer-Uniscale [14-15] (appendix 2); and direct utility-based scaling techniques such as the standard gamble and time trade-off. In contrast, a specific question focuses on a specific aspect of HRQL such as an element of physical function, psychological well-being, or social function; or a specific symptom of disease or side effect of treatment such as pain or hair loss. Specific questions are often grouped together by domains in a HRQL instrument, providing a comprehensive profile-based assessment across physical, psychological and social dimensions. Examples of these instruments are described in section 2.3.3 below. Global and specific questions have different uses. A global question can provide an estimate of overall HRQL at one point in time, changes in overall HRQL over time, and the difference in overall

HRQL between groups such as those allocated to differing treatments in a clinical trial. A profile of specific questions can provide descriptive data about the specific aspects in which the impairment of deterioration occurred [1].

Another way of classifying HRQL measures is by whether the scaling method used to assign numbers to the responses on the question is value-based or utility-based. A value-based scaling method is one that expresses a respondent's perceptions about the presence, intensity or severity of a symptom, function or disability. In contrast, a utility-based scaling method is one that expresses a respondent's strength of preference for a particular outcome or health state [5]. The value-based and utility-based scaling methods have different uses. Value-based scaling methods are commonly used to evaluate and compare the effects of diseases and treatments on various dimensions of HRQL in clinical trials and surveys. Utility-based scaling methods are commonly used to inform choices between alternate therapies by decision-makers. This is because utility-based scaling methods capture both the person's preference, and their attitude towards risk for future outcomes that are uncertain; and express it in a standardised way that can be integrated with other data about the probabilities and values of outcomes using econometric techniques and metrics such as quality-adjusted life-years [5, 9, 16-17]. This is discussed in more detail in section 2.4.

The next section discusses methods for evaluating the effects of diseases and treatments on HRQL with value-based scaling methods.

2.3 Measurement of HRQL with value-based scaling methods

2.3.1 Conceptual framework

A value-based scaling method is one that expresses a respondent's perceptions about the presence, and intensity or severity of a symptom, function or disability [5]. Unlike utility-based scaling methods discussed in the next section, the focus of value-based scaling methods is on current health rather than future outcomes. The value-based scaling method is sometimes referred to as the psychometric scaling method because the way that numerical scores are assigned to the subjective responses comes from the psychometric tradition. Psychometric techniques enable perceptions such as the statement 'I feel severe pain', which is not inherently quantitative, to be converted to a level on a response scale. Examples of response scales are binary scales: for example pain 'present' or 'absent'; ordinal scales representing increasing severity: for example 3 for 'mild' pain, 4 for 'moderate', 5 for 'severe'; or a visual analogue scale where a respondent marks a point on a line anchored from 'no pain' on the left and 'the most severe pain' on the right [5].

Value-based scaling methods can express a respondent's perceptions about global HRQL or health status with a single-item global scale, or express a respondent's perceptions about specific aspects of HRQL. The merits of each approach are discussed in sections 2.3.2 and 2.3.3 below.

2.3.2 Single-item global scales

A single-item global scale is one that asks respondents for a unified assessment of their global health status or HRQL. The main strength of a single-item global scale is that it is easier to interpret than a profile of multiple items about specific aspects of HRQL. A single rating can help to determine the net difference in overall HRQL between groups or changes over time. Another advantage is that a single-item global scale is quick and easy to elicit, which reduces the burden on patients and staff. It also reduces the burden on statisticians and readers by providing data that is simpler to analyse and report [18].

The main limitation of a single-item global scale is that it is less reliable and informative than a profile of multiple items. Single-item global scales elicit

responses that are open to interpretation by each respondent and, when used alone, do not provide information about what aspects of quality of life are important to each respondent. This is beneficial, in that it allows each respondent to focus on what is most important to them in rating global quality of life, even if their conceptualisation of quality of life is different to that of other patients. It does however pose additional problems because global ratings are more vulnerable to certain types of bias and measurement error that make them less reliable than a profile of multiple items. Response shift refers to respondents changing their conceptualisation of quality of life over time and therefore effectively answering a different question on each occasion. For example, as subjects deteriorate, their expectations may also decrease, and their responses do not differ as much as would be expected. Another type of bias is called context bias. Subjects who are older or with worse disease may assign similar responses to that of subjects who are younger or with less advanced disease. End-aversion bias means that subjects rarely assign responses at the higher or lower extreme of a scale. Ceiling and floor effects occur where subjects assign a similar response despite different HRQL, because the scale has a restricted range of options [5, 18-19]. Single-item global scales compared to indices derived from a profile of multiple items are more susceptible to these types of bias, because the random error in multiple items tends to cancel out (see section 2.5.2) [18].

Another problem with single-item global scales is that respondents may find that the elicitation task is conceptually difficult, which may lead to imprecise responses. For a subject to rate their HRQL with a global scale, they must consider all aspects of their HRQL, implicitly adding additional weight to the specific aspects of HRQL that are most important to them, and ignoring irrelevant aspects. The biases and imprecision of single-item global scales reduce their power to detect small but meaningful differences in HRQL between treatment groups [5, 20-22]. Instruments containing items about specific aspects of HRQL are often used either to substitute for, or to complement a single-item global scale, in an attempt to overcome some of these problems [18].

2.3.3 Profile-based instruments

Multiple items about specific aspects of HRQL are often grouped together in an instrument that provides a profile across a range of HRQL dimensions. These instruments are sometimes referred to as 'profile-based instruments' or 'Health status assessments'. They are designed to compare levels of functioning in specific dimensions of HRQL between groups, and changes in their function over time [2, 5, 17].

Profile-based instruments vary according to the type of questions that they contain. They can be classified as generic, disease-specific or domain-specific. Generic instruments assess aspects of HRQL that are applicable to a wide range of populations and interventions. Disease-specific instruments cover all dimensions of HRQL but focus on particular aspects that are relevant to a specific population: for example patients with advanced cancer or diabetes. Domain-specific instruments do not cover all dimensions of HRQL, but instead focus on a particular dimension of HRQL such as emotional function, or a specific symptom or side effect of a disease or treatment [2, 5].

The multiple items contained in a generic profile-based instrument will typically address key dimensions of HRQL such as physical function, psychological and emotional well-being, and social function. These dimensions are applicable to patients with a range of diseases and treatments, and to the general population. The advantage of generic instruments over disease-specific instruments is that they allow straightforward comparisons between different populations. They are useful for monitoring patients with multiple diseases, for comparing the health status of patients with different diseases, and for comparing patients with members of the general population [5].

The items contained in a disease-specific profile-based instrument are more pertinent to a specific disease or treatment. They include questions about specific symptoms and side effects that are likely to be encountered by patients with a particular disease as well as questions about general aspects of HRQL. For example, nausea and vomiting are commonly experienced by patients with cancer, due to the effects of

cancer and chemotherapy. A generic instrument may lack any items about nausea and vomiting, but they will be included in most disease-specific instruments for cancer. The inclusion of disease-specific items in a disease-specific instrument should improve its responsiveness to detect changes in disease-specific aspects of quality of life that are not addressed by a generic instrument. The major limitation of disease-specific instruments is that they may hamper comparisons between populations with different diseases because of differences in the items that are included [2, 5].

Examples of a generic and three commonly used disease-specific instruments for cancer are shown in table 2.1. The Medical Outcomes Study 36-item Short-Form Survey (SF-36) include 35 questions about generic aspects of HRQL that are grouped into 8 dimensions, as well as a transition question that asks how the respondents health has changed compared to one year ago [23]. The disease-specific instruments in table 2.1 include some questions about generic aspects of HRQL, in addition to the specific symptoms that are relevant to cancer patients.

The purpose of a study will determine the most suitable profile-based instrument. In a clinical trial, a disease-specific instrument will often be used because it is more likely to detect differences between treatment groups and changes in HRQL over time. A generic instrument may be used in combination with a disease-specific instrument if the researcher wishes to study broader aspects of HRQL, but care must be taken not to unreasonably increase patient burden [2, 11].

Both types of profile-based instruments often contain a large number of items, and generate a large number of responses. The next section describes methods for reducing the multidimensionality of the data to aid analysis and interpretation.

Table 2.1 Generic and cancer-specific profile-based HRQL instruments

Instrument	Type	Number of items	Dimensions (Number of items)	Region of development
SF-36	Generic	36	Physical functioning (10) Role limitations due to physical health problems (4) Bodily pain (2) Social functioning (2) General mental health (5) Role limitations due to emotional problems (3) Vitality, energy or fatigue (4) General health perceptions (5) *Transition question (1)*	USA
EORTC QLQ-C30	Cancer-specific	30	Physical function (5) Role function (2) Cognitive function (2) Emotional function (4) Social function (2) Fatigue (3) Nausea (2) Pain (2) Symptom scales (6) Global quality of life (2)	Europe
FACT-G	Cancer-specific	27	Physical well-being (7) Social/family well-being (7) Emotional well-being (6) Functional well-being (7)	USA
UBQ-C	Cancer-specific	31	Physical function (3) Distresses due to physical and psychological symptoms (21) Social/usual activities (4) Self-care (1) General health (1) Global health status (1)	Australia

SF-36, Medical Outcomes Study 36-item Short-Form [23]. EORTC QLQ-C30, European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire-Cancer [24-25]. FACT-G, Functional Assessment of Cancer Therapy – General [26]. UBQ-C, Utility-Based Questionnaire-Cancer [12-13].

2.3.4 Scoring of profile-based instruments

Profile-based instruments typically include multiple items across several dimensions of HRQL. Analysing data about multiple items can pose methodological problems concerning multiple testing and is difficult to interpret. To overcome these problems, some instruments use a scoring system to aggregate the individual items into a smaller number of subscales that represent each dimension [17, 27]. In this section, scoring systems that derive subscales from multiple items are discussed. In section 2.5, scoring systems that derive a utility index from multiple items are discussed. Both approaches are relevant to the work presented in this thesis.

The standard approach to deriving subscales from individual items is referred to as the 'equally-weighted' approach. It derives a subscale by combining the simple average of the responses on each related item. Related items may be selected by expert opinion, or by psychometric techniques such as factor analysis [28]. The result is expressed in a standardised form, such as a percentage of the maximum score achievable for that dimension [2, 4-5, 29]. For example, the generic SF-36 instrument (referred to in the previous section) consists of 36 items. Its scoring algorithm produces a profile of subscales across eight dimensions about physical functioning, physical role, bodily pain, general health, vitality, social functioning, emotional role and mental health; by applying equal weight to each component item. The first four dimensions contribute greatest weight to a Physical Component summary score, and the remaining four contribute greatest weight to a Mental Component Summary score. However the weights were generated by factor analysis [23]. The cancer-specific EORTC QLQ-C30 instrument consists of 30 items. The scoring algorithm produces a profile of subscales across 8 dimensions about physical function, role function, cognitive function, emotional function, social function, fatigue, nausea, pain, and global quality of life [25]. For both instruments, the subscales are the simple averages of the items. The result is linearly transformed to a scale from 0 to 100 with a higher score representing a higher level of function [5].

Aggregating multiple items within one dimension into a subscale using the equally-weighted approach is often used because it aids interpretation, is straightforward, and

gives similar results to more complex methods that assign different weights to each item based on their importance [27, 29-30].

Aggregating multiple items *across* dimensions into a single index of overall HRQL using the equally-weighted approach is appealing but can be problematic. It is intended to express a unified assessment of the impact of a disease and its treatment on daily life that is more precise and informative than a single-item global scale because of its better psychometric properties [31] (discussed in sections 2.3.2 and 2.3.3). However, the problem with assigning equal weights to unrelated dimensions of HRQL is that it assumes that each dimension is equally important to patients, and that questions are included about all important aspects of HRQL. Some studies have shown that indices calculated with equal weightings can give results that are inconsistent with patients' views as reflected in a global measurement, particularly if the index is calculated from less than 40 items [28, 32]. Biased results could lead to incorrect conclusions in analysis of treatment effects [11, 32-33].

Indices of overall HRQL derived by assigning equal weights to items across dimensions have been developed for cancer-specific instruments including the Functional Living Index-Cancer (FLIC) (referred to as the 'overall FLIC score')[34], and the Functional Assessment of Cancer Therapy (FACT) (referred to as the 'FACT-G total score')[26]. However this approach is generally not recommended [32, 35], especially for primary analysis, because of the potential for biased results that was discussed in the previous paragraph [29]. Better approaches that derive an index of overall HRQL by giving weights to each item or subscale that reflects their relative importance are described in section 2.5.

2.3.5 Applications and limitations

Value-based scaling methods provide useful descriptive information about the effects of diseases and treatments on specific aspects of HRQL and of global HRQL. However they do not express the overall desirability of health states in a way that can be directly integrated with other data about the probabilities and values of outcomes to inform decision-making. The next section discusses direct utility-based scaling methods that go beyond the descriptive information obtained with value-based scaling methods.

2.4 Measurement of HRQL with direct utility-based scaling methods

The utility-based scaling method, sometimes referred to as the preference-based method, is one that provides an estimate of the overall desirability of a health state expressed on a scale from zero, representing death, to one, representing perfect health. The utility-based scaling method arose from the econometric tradition. A utility is a quantitative expression of an individual's preference for a particular health state under conditions of uncertainty. Utilities have special properties that allow their integration with other information about the probabilities and value of outcomes using the econometric techniques of decision analysis and cost-utility analysis based on the quality-adjusted life-year (QALY) approach [5, 9, 16, 36]. This facilitates their use to inform decision-making.

2.4.1 Conceptual framework

The conceptual framework of the utility-based scaling method is best understood by contrasting the differences between utilities and values. The traditional interpretation of von Neumann-Morgenstern utility theory states that utilities and values are related concepts that differ mainly in the conditions under which the judgments are made. Utilities are numbers that represent the strength of an individual's preferences for different health states under conditions of uncertainty, while values are the numbers that people assign to different health states that are certain. In other words, the utility-based scaling method expresses preference for outcomes that are uncertain, whereas the value-based scaling method reflects preference for outcomes that are certain. Utilities and values both reflect an individual's level of satisfaction, distress or desirability for a particular health state. In this interpretation, the main difference between utilities and values is that utilities incorporate a respondent's attitude to risk [17, 37-39].

The conceptual framework underlying utilities has been used in health economic applications since the 1940s. The von Neumann-Morgenstern utility theory, also known as the theory of rational decision making under uncertainty, is an extension of the utility theory of economics. It was developed as a model of how a rational individual ought to make decisions when faced with uncertain outcomes [40].

Fundamental axioms of choice underlie the von Neumann-Morgenstern utility theory, as described by Torrance et al [40]. The first axiom states that respondents will have preferences for one outcome compared to another, and that these preferences are transitive. This means that if two potential outcomes exist, o and o'; either o will be preferred to o', or o' to o, or the subject will be indifferent between o and o'. The preferences are transitive because if o is preferred to o' and o' is preferred to o'', then o is preferred to o''. The second axiom states that a rational individual will be indifferent between a one-stage and a two-stage gamble. This will be illustrated by description of the standard gamble method of eliciting utilities in the next section. The third axiom is of continuity of preferences. This implies that if an individual prefers an outcome o to o', and o' to o''; then there is a probability p at which the individual is indifferent between the certain outcome of o', and a gamble between o with probability p, and o'' with probability (1-p) [40]. These axioms of utility theory determine how utility-based scaling methods such as the standard gamble are used to express the preferences of respondents for differing health states, and how utilities about the desirability of outcomes are combined with information about the probabilities of outcomes using econometric techniques such as decision analysis.

Alternate interpretations of utility theory exist. Richardson states that decision making under uncertainty is not essential to obtain utilities [41]. He also argues that utility-based scaling methods based on decisions under certainty (as discussed in the next section) have both theoretical and empirical advantages for generating QALYs [41].

2.4.2 Direct utility-based scaling methods

There are two main approaches for eliciting utilities for health states, which are 'direct' and 'indirect'. The direct approach uses a utility-based scaling method, and is discussed in this section. The indirect approach converts responses elicited with a value-based scaling method into a utility index by applying a scoring algorithm, and is described in section 2.5.

The scaling method is the specific task required of a respondent to assign their strength of preference to a health state. There are several direct utility-based scaling

methods and controversy exists as to which method is best. The most common methods are the standard gamble and time trade-off. The rating scale is related but uses a value-based scaling method rather than a utility-based scaling method [9, 16, 42]. This section describes the specific task required of respondents for each method, and discusses the relative merits of each approach.

Standard gamble

The standard gamble method requires respondents to choose between uncertain outcomes that may occur in the future, as described by Drummond [9], Torrance [40] and Froberg [42]. The respondent is given a choice between accepting a defined health state with certainty; or taking a gamble between a treatment that may give a better outcome (such as perfect health) with a probability p, or a worse outcome (such as immediate death) with a probability of (1-p). The probability p is varied until the respondent is equally willing to accept (ie. indifferent to) the defined health state with a certain outcome and the gamble. This value of p represents the utility of the health state. For example, a respondent is asked to consider the desirability of a defined health state relative to perfect health or death. The respondent may prefer a gamble with a 99% chance of returning to perfect health and a 1% chance of immediate death, rather than remaining within that defined health state for a fixed period of time. However they would prefer the defined health state to a gamble with a 50% chance of returning to perfect health and 50% chance of immediate death. They are indifferent to a gamble with an 80% chance of returning to perfect health and a 20% chance of immediate death versus remaining in the defined health state. The utility of the defined health state is then 80% or 0.8. The method can be varied as to the duration of the health state (such as 5 years), and the outcome of the gamble alternative (usually perfect health and death) [9, 43]. The standard gamble is generally administered by a trained interviewer [9, 40, 42].

Time trade-off

Like the standard gamble, the time trade-off requires respondents to choose between outcomes that occur in the future, however the outcomes occur with certainty rather than uncertainty. The time trade-off was developed by Torrance et al for use in health research as an easier alternative to the standard gamble [44]. As described by Drummond, Froberg and Feeny [9, 42, 45], the time trade-off method involves respondents being given a choice between a longer period of time (t) in a defined

health state with less than perfect health, or a shorter period of time (x) with perfect health. The period of time with perfect health (x) is varied until the point of indifference between the two health states: this reflects how much time a respondent is willing to trade-off in order to have better health. The utility of the health state is x/t. For example, a respondent is asked to consider the desirability of a defined health state. The respondent may prefer to live for 4.5 years in perfect health rather than 5 years in the defined health state. However they would prefer 5 years in the defined health state rather than 1 year with perfect health. They are indifferent to 4 years with perfect health versus 5 years in the defined health state. The utility of the defined health state is then 4/5 or 0.8. The method can be varied as to the duration of the health state t (such as 5 years), and the outcome of the health state alternative (usually perfect health) [9, 43]. Like the standard gamble, the time trade-off is usually administered by a trained interviewer [9, 42, 45].

Rating scale

The rating scale, also referred to as the visual analogue scale, involves respondents rating the desirability of a defined health state by placing it at some point on a line, anchored by clearly defined endpoints which are conventionally 'death' and 'perfect health'. If respondents are asked to rate more than one health state, then they are asked to place the health states so that they reflect the rank order of the states, and the intervals between the placements reflect the perceived differences between the health states. The line may vary in length, be vertical or horizontal, and may have intervals marked out with different values. The rating scale is usually self-administered [9, 16, 42].

Comparison of methods

A number of reviews with differing conclusions highlight the controversy regarding the optimal direct utility-based scaling method. Different authors favour the standard gamble, time trade-off, or rating scale, because of their underlying theory, reliability, validity, or ease of use [16, 42, 46-50].

From the theoretical perspective, the standard gamble is the criterion method for eliciting utilities. It is directly founded in von Neumann-Morgenstern utility theory because it measures preferences for outcomes under uncertainty. Unlike the standard gamble, the time trade-off does not measures preferences for outcomes under

uncertainty, because there are no probabilities in the time trade-off question. It does however take into account a respondents' attitude to choice, and also tests preference for immediate versus delayed outcomes (ie. time preference). It is commonly used as a substitute for the standard gamble because it was designed to give comparable scores. The rating scale does not measure preferences for outcomes under uncertainty, because the responses are assigned by a value-based scaling method. Therefore the scores elicited with a rating scale are values rather than true utilities [9, 16, 51-52].

Both the standard gamble and the time trade-off have been shown to have acceptable validity, reliability and responsiveness in a wide variety of contexts [16]. The main limitation of these methods is the difficult cognitive task required of respondents, leading to a larger number of refusals, missing values and inconsistent responses than other methods [45]. Utilities derived from the standard gamble are susceptible to risk aversion, leading to inflated values [42]. The time trade-off was developed to overcome the difficulties of explaining probabilities to patients in the standard gamble. It is easier to administer than the standard gamble, but is still cognitively demanding [16]. Utilities derived from the time trade-off tend to be lower than those derived from the standard gamble [49]. Greater inconsistencies in rating of patient preferences have been identified with the time trade-off than with other scaling methods – for example rating all health states as equal or illogical ordering of utilities - have been identified, particularly in older and less educated individuals, as well as those with cognitive impairment or poorer health status [47]. Another limitation of the standard gamble and time trade-off is that some respondents are unwilling to trade-off or risk any of their remaining life-expectancy to improve their health state, which leads to suboptimal health states having the same utility as perfect health [16].

The rating scale has a high rate of completion and reliability, given the easier cognitive task demanded of respondents, and is less costly to administer [16]. One limitation of the rating scale is 'end-of-scale aversion', where respondents avoid putting states very close to the most and least desirable ends of the scale [45]. The greater problem is that the responses on the rating scale are expressed with a value-based rather than utility-based scaling method, so do not provide utilities. This is

problematic, because they do not measure HRQL in a way that can be combined with quantity of life to generate QALYs. However unlike profile-based instruments, global scales require respondents to implicitly add additional weight to specific aspects of HRQL that are more important (section 2.3.2). This is why methods exist to map the obtained value to a utility using a transformation function [16], at least at a population level (see chapter 5, section 5.6 for more discussion). In this sense, a global rating scale is a hybrid between a value-based and a utility-based scale.

In summary, there is no single direct utility-based scaling method that is optimal for all situations. When selecting the scaling method for a particular study, it is important to assess the purpose of the study and the importance of deriving a true utility estimate, the characteristics of the respondents who will have to perform the scaling method, the framing of assessment, and the resources available. Regardless of the scaling method, great care must be taken when eliciting utilities from respondents, to elicit valid and reliable results [40].

2.4.3 Perspectives for utilities

Utilities can be elicited from lay people, health care professionals, relatives or patients. Lay people can only assign utilities about their preference or desire for hypothetical health states. This is referred to as 'decision utility', and is based on the concept of 'wantability'. Patients can assign utilities about hypothetical health states, health states they have previously experienced, or their current experience of a health state. This is referred to as 'experienced utility' and is based on the concept of 'hedonic' experience [53]. The distinction is important because marked differences in valuations between different groups have been reported [54-58]. This difference in perspectives may have significant implications for clinical and economic decisions that incorporate utilities and QALYs [56-58]. Because utilities assigned by patients are typically higher than utilities assigned by lay people, their use is likely to redirect priorities away from treatments that improve HRQL and towards treatments that extend life [59]. For example, in the extreme situation where a patient has complete adaptation to a health state and assigns a utility of 1, there is no gain from a treatment that improves HRQL. These effects are described in detail in the discussion chapter of this thesis (Section 8.5.1).

For example, a patient typically assigns a higher utility to a health state than a lay person [55-58]. This may reflect partly the lay person's difficulty appreciating what a hypothetical health state is really like, and partly the patient's adaptation to their own health state [16, 58, 60-62]. Dolan notes that lay people overestimate the losses associated with transition to health states because of focus on the immediate impact which may lessen over time, and focus on a specific health domain rather than unaffected health domains. Dolan also notes that patients' adaptation to health states is common but not universal [53]. It has been recognised that the preferences and attitudes of patients in different clinical contexts may differ, because patients with different diagnoses, stages of disease and treatments may assign different importance to different aspects of HRQL [32, 63-65]. It has also been recognised that the preferences and attitudes of lay people in different countries may differ, because of differences in demographic background, social and cultural values, and political and economic systems [66-67].

Controversy exists about the suitability of utilities that are valued from the perspective of lay people versus patients [58-59, 68]. The choice of perspective for eliciting utilities should reflect the viewpoint from which the results will be interpreted [16, 61]. Health economic guidelines generally recommend the use of generic utility-based instruments based on the perspective of lay people [69-71]. The main argument for using the perspective of lay people for informing funding and policy decisions is that the primary objective in a publicly funded health system is to maximise health for society [9]. A limitation of using general population samples is that their assessments are less well informed, and limited to the supplied descriptions [72]. It is generally recommended that the perspective of patients is used to inform clinical decision making [62, 73]. The main argument for using the perspective of patients for clinical decisions is that the primary objective is to maximise health for the individual patient experiencing that condition. Some also argue that the perspective of patients should be used to inform funding and policy decisions, because patients better understand what it is like to live with a particular disease [56, 59], but this argument is controversial because it runs counter to prevailing health economic theory and guidelines [69-71].

In summary, it is important to recognise that: utilities are dependent on the experiences, attitudes and beliefs of the respondents; and a judging population should be chosen to be appropriate for the research or policy question that is being answered [16, 54].

2.4.4 Applications and limitations

Utilities elicited by direct utility-based scaling methods are available for a wide range of diseases [74]. Utilities and the QALYs that are generated from them are a useful way to compare treatments for cancer and other diseases, because they can be evaluated on a common metric that incorporates disparate treatment effects. The utility combines the improvements in HRQL due to relief of disease symptoms, and the deteriorations in HRQL due to treatment-related side effects [9, 75]. The QALY approach combines the net effects of treatments on HRQL with the effects on survival. Analyses of cancer trials in terms of utilities and QALYs are increasingly used to inform economic decisions about cancer treatments [76-83], but can also be used to inform clinical decisions [60, 84-89].

Although direct utility-based scaling methods are a standard way of eliciting utilities from respondents, they have several limitations. One is that the task is complex, resource-intensive, and can be distressing or burdensome if patients are required to assign utilities for their own health state [16, 32]. Another limitation is that the resultant utility scores do not provide descriptive information about the dimensions of HRQL that influence the utility [38]. For example, in the interpretation of a clinical trial, a treatment may improve overall HRQL as reflected by a greater utility, however this improvement could reflect large positive improvements in some aspects of HRQL, at a cost of detriments in other aspects. Therefore the single direct utility-based scaling method may obscure important trade-offs.

Because of these limitations, a more practical approach for obtaining utility scores in clinical trials is to derive them indirectly from a utility-based instrument. This approach is discussed in the next section.

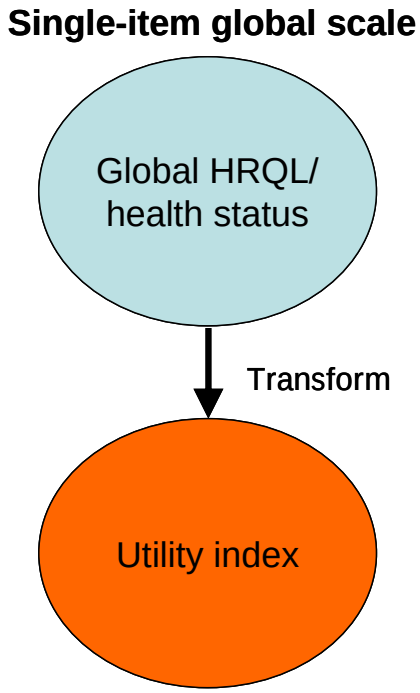
2.5 Measurement of HRQL with utility-based instruments

An alternative and practical approach to obtain utilities for health states is to use a utility-based instrument. A utility-based instrument uses a scoring algorithm to convert the responses from a questionnaire that elicits ratings about various dimensions of HRQL, into a single index that is expressed with a utility-based scaling method. The scoring algorithms are valued in surveys, where a sample of respondents is asked to assign utilities to the health states defined by the questionnaire. Patients then complete the questionnaire during a clinical trial or survey, and based on their ratings are assigned to a discrete health state category. The health state category is then mapped to a pre-existing utility score by applying the scoring algorithm. Utility-based instruments vary in the type of items contained within the questionnaire (single-item global scale, multiple generic items, or multiple disease-specific items), and the perspective from which they are valued (typically of lay people or patients) [9, 38, 42, 90].

2.5.1 Conceptual framework

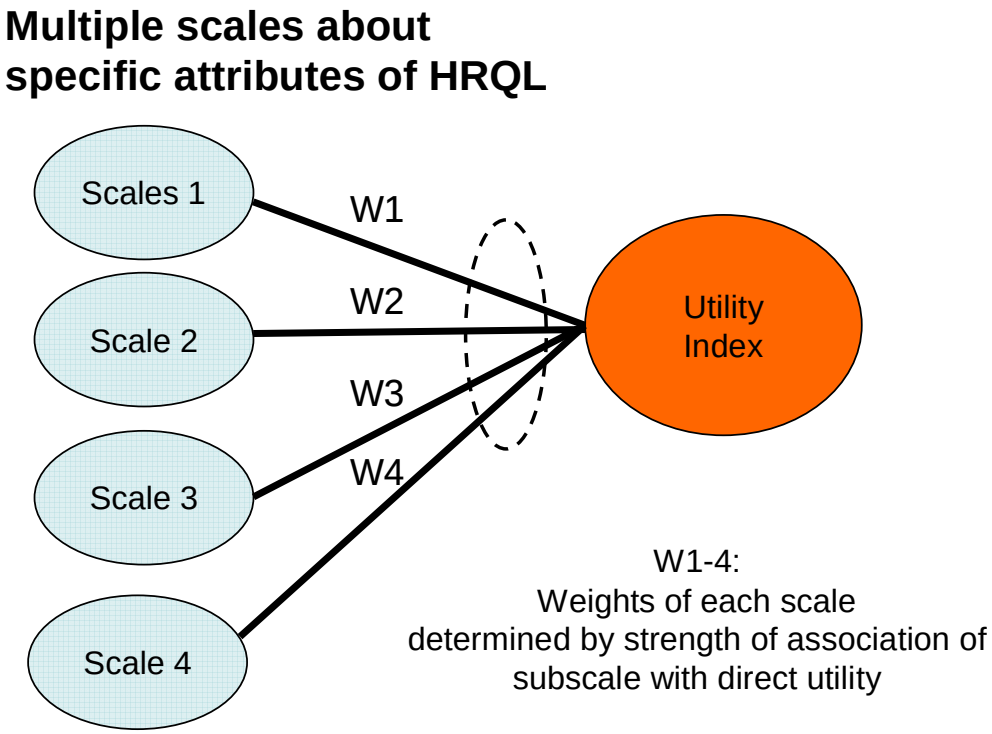
The broad conceptual framework for a utility-based instrument is that an individual's utility for a given health state can be determined by their perceived health status and quality of life in that health state. There are two distinct but related approaches and conceptual frameworks which will be referred to as the 'global health preference' approach and the 'multi-attribute health preference' approach [9, 91] (figures 2.1 and 2.2).

Figure 2.1 Deriving a utility index with the global health preference approach



HRQL, health-related quality of life.

Figure 2.2 Deriving a utility index with the multi-attribute health preference approach



A utility-based instrument based on the global health preference approach comprises a single-item global scale such as a rating scale about global quality of life or health status. It is important that the scale has the anchors of full health and death to correspond with the anchors of a utility scale. The utility scores are derived by applying a scoring algorithm to the global scale such as a mathematical power transformation [92] (figure 2.1).

The underlying conceptual framework for the global health preference approach is that a value-based scaling method, such as a global rating scale, measures preferences under certainty, and a utility-based scaling method measures preferences under uncertainty. The link between the value and utility is an individual's attitude to trading length and quality of life and to risk, which is estimated by the mathematical transformation [37, 92-95].

Other utility-based instruments are based on the multi-attribute health preference approach. They are comprised of multiple items about various dimensions of HRQL. The utility scores are derived by applying a scoring algorithm that applies weights to the responses on each item that reflects its relative importance. For example, greater weight may be applied to alterations in physical function compared to alterations in appearance. The underlying conceptual framework is that an individual's utility for a given health state can be determined by their perceived status in multiple dimensions of health and quality of life, such as physical, emotional, and social dimensions [9, 39, 90].

The merits of a utility-based instrument comprised of a single-item global scale and based on the global health preference framework, versus one comprised of multiple items and based on the multi-attribute health preference framework, is discussed in the next section.

2.5.2 Utility-based instruments comprised of a global scale or multiple items

There are strengths and weaknesses of utility-based instruments that are comprised of a single-item global scale, or of multiple items about specific aspects of HRQL. As discussed in section 2.3.2, the major strengths of single-item global scales are that

they are simpler to elicit and analyse compared to a set of multiple items, and allow respondents to assign a unified assessment of their global HRQL. The major limitation is that they are less reliable and informative than an index derived from multiple items.

A utility-based instrument comprised of multiple items has several advantages over a utility-based instrument comprised of a single-item global scale. The multi-dimensionality of the data gives more descriptive information about HRQL than a single-item global scale. An instrument based on multiple items also has better measurement properties than one based on a single-item. This finding is explained by classical test theory. An index derived from multiple items will provide a more reliable estimate than a single-item scale, because of reductions in random error and bias afforded by averaging. An index is therefore more stable and reliable than one based on a single-item scale [22, 31, 96].

One potential limitation of a utility-based instrument comprised of multiple items is its dependence on the inclusion of all important aspects of HRQL. To be valid, the instrument should include scales about all important aspects of HRQL, including symptoms due to disease and side effects of treatment. If important items are omitted, then the instrument may miss important differences in HRQL between groups [32, 97-98].

One complexity of a utility-based instrument comprised of multiple items is its reliance on the scoring algorithm. The scoring algorithm should optimally weight the items to reflect the relative importance of each aspect to the population that the researcher is trying to reflect in the decision-making [9, 16, 99]. Evidence suggests that different populations assign very different weights to different aspects [64]. The preferences and attitudes of lay people in different countries differ, perhaps because of differences in demographic background, social and cultural values, and political and economic systems [66-67]. It has also been shown that the preferences and attitudes of cancer patients in different clinical contexts differ, perhaps because patients with different cancer diagnoses, stages of disease and treatments assign different importance to different aspects of HRQL [32, 63-65]. If an instrument lacks important aspects (as stated above), or is inappropriately weighted, then it may give a

biased estimate of a treatment effect. Once a standard instrument comprised of multiple items and its scoring algorithm is developed and validated, modification by adding relevant items, dropping unnecessary items, or altering weights for a particular purpose, all necessitate repetition of the development and validation process [32]. For example, significant work has been required to generate new country-specific scoring algorithms for utility-based instruments, where the attitudes of lay people in different countries are thought to differ [66, 100-102].

In summary, utility-based instruments comprised of only a single-item global scale are more transparent, and are simpler to administer, analyse and interpret, because there is only one scale. But they are less reliable and informative than multi-item scales. Utility-based instruments comprised of multiple items have better measurement properties, but are less transparent, and are only valid if all relevant aspects of HRQL are included and are appropriately weighted. In the next section, the merits of multi-item utility-based instruments comprised of generic or disease-specific items are discussed.

2.5.3 Multi-item instruments containing generic or disease-specific items

Utility-based instruments comprised of multiple items can be generic, in that they only contain items about generic aspects of HRQL, or disease-specific, in that they contain items about aspects of HRQL that are particularly relevant to a specific disease. Generic instruments like the EuroQol EQ-5D [103-105], Health Utilities Index (HUI3) [106] or SF-6D [107] ask about core aspects of HRQL that are of interest in a wide range of settings. The main argument for using a generic utility-based instrument is that it allows comparisons across a wide range of diseases and healthy populations [68, 108-109]. However a generic instrument is likely to provide an inadequate description of many diseases, so the utility scores that it generates may be insensitive to differences between individuals with that disease [108, 110-112]. More recently, disease-specific utility-based instruments have been developed that ask about specific aspects of HRQL relevant to that disease or condition [97-98, 110, 113]. The main advantage of a disease-specific instrument over a generic instrument for generating utility scores is that it should be more sensitive to differences in HRQL between individuals with a particular condition, such as cancer [95, 98, 110] or a range of other diseases [97, 111, 113-114]. Another advantage of using a disease-specific utility-based instrument is that it provides data on specific aspects of

HRQL, overall HRQL, and utility with a single questionnaire and increases the availability of utility data for comparisons of treatment from randomized clinical trials [98].

The major limitation of using disease-specific, utility-based instruments is that the utility scores they provide may not be comparable to those derived from other instruments, particularly generic instruments, because the dimensions of health status and HRQL that they cover are different [68, 108-109]. For this reason, disease-specific instruments are best suited to treatment comparisons within a particular disease used to inform clinical decisions. In this context comparisons across other diseases and healthy populations are less important, but coverage of aspects relevant to the patients under study is crucial. Others have argued that disease-specific instruments may also be suitable for treatment comparisons across all diseases to inform health funding and policy decisions if the scoring algorithm is derived using a valuation technique and population sample that is similar to a generic instrument, and the utility scores are shown to be comparable [97].

2.5.4 Perspectives

Utility-based instruments also vary by the perspective from which they are valued. The perspective is determined by the characteristics of the judges who assigned utilities to the health states that were used to derive the scoring algorithm. Most utility-based instruments are based on the perspective of lay people, but utility-based instruments have also been developed that are based on the perspective of patients [94, 97-98]. In section 2.4.3 it was noted that the appropriate perspective is determined by the context in which the utilities are to be applied. The perspective of lay people is best suited to informing economic decisions, and the perspective of patients is best suited to informing clinical decisions.

2.5.5 Applications and limitations

Utility-based instruments are commonly used in clinical trials for cancer and a range of other conditions [9, 115]. They provide descriptive information about the effects of disease and treatment on HRQL, and also provide utility scores that can be used to generate QALYs. The main potential limitations of utility-based instruments are that they may be comprised of inappropriate items, or use a scoring algorithm that is based on an inappropriate population. The next section describes standard

approaches for selecting appropriate items, and deriving appropriate scoring algorithms for utility-based instruments.

2.6 Deriving scoring algorithms for utility-based instruments

A key component of utility-based instruments is the scoring algorithm that converts ratings on the questionnaire into a utility index. In this section, considerable detail will be given to the approaches for deriving a scoring algorithm, because this is the focus of the work presented in this thesis. The task of deriving the scoring algorithm takes place in three stages. The first stage is to determine the items and response options that will comprise the questionnaire of the utility-based instrument. Instrument developers refer to this process as developing a 'health state classification system' that describes a series of health states in terms of levels of impairment on one or 'attributes'. The second stage is to perform a 'valuation survey', where the health states described by the health state classification system (ie. all the health states described by all the conceivable combinations of responses on each item of the questionnaire) are valued. The third stage is to produce a scoring algorithm that assigns a utility to every conceivable health state described by the health state classification system. Once the scoring algorithm is produced, the utility-based instrument can be used in a clinical trial or other situation. Respondents complete the questionnaire of the utility-based instrument that describes their level of impairment on each attribute in the health state classification system. The scoring algorithm is then applied to map the respondent's health state, as represented by the ratings to each item on the questionnaire, to a utility between zero and one that represents the desirability of their health state [9, 39, 90].

2.6.1 Determining the items and response options

The first stage in deriving a scoring algorithm is to determine the items and response options that will describe a series of health states in terms of level of impairment on each item. Each item can have two or more levels ranging from best to worst, such as 'No impairment', 'Moderate impairment' and 'Severe impairment'; or less commonly a numerical scale which could range from 0 to 10. The simplest instrument has only a single-item global scale, and is based on the global health preference framework referred to in section 2.5.1. A multi-attribute instrument has multiple items relating to relevant dimensions of HRQL such as physical,

psychological and social function, and is based on the multi-attribute health preference framework referred to above.

Most utility-based instruments limit the number of items and their levels to reduce the size and complexity of the instrument. One reason for limiting the number of items is that a judge rating health states in a valuation survey (as discussed in the next section) may be unable to reliably process information on more than five to nine attributes. Another reason is that producing a scoring algorithm from a large number of items is mathematically complex because of the large number of possible interactions between items [39, 116].

Items and response options for commonly used utility-based instruments are shown in table 2.2. For example, the Disability and Distress Scale is one of the earliest and simplest multi-attribute utility-based instruments. Its health state classification system consists of two items: distress (4 levels) and disability (8 levels), describing 32 health states (4*8). 29 of the health states are plausible and 3 are implausible (eg. chair-bound but no distress) [117-118]. The EQ-5D utility-based instrument has five items, each with three levels, and describes 243 health states [103-105]. The SF-6D utility-based instrument has 6 items, each with 4-8 levels, and describes 18000 health states [107].

Table 2.2 Comparison of scoring algorithms for utility-based instruments

	Subjective health estimation scale	Disability and Distress scale	EQ-5D	HUI3	SF-6D
Items and response options					
Number of items	1	2	5	8	6
Item description (and number of response options)	Health status (100)	Disability (8) Distress (4)	Mobility (3) Self-care (3) Usual activities (3) Pain (3) Anxiety/depression (3)	Vision (6) Hearing (6) Speech (5) Ambulation (6) Dexterity (6) Emotion (5) Cognition (6) Pain (5)	Physical functioning (6) Role limitation (4) Social functioning (5) Mental health (5) Bodily pain (6) Vitality (5)
Potential number of health states	1*100 =100	4*8 – 3 = 29	$3^5 = 243$	6*6*5*6*6*5*6*5 = 972000	6*4*5*5*6*5 = 18000
Valuation survey					
Scaling method	TTO	Magnitude estimation	TTO	Rating scale/ Standard gamble	Standard gamble
Population	Patients	Health care workers	Lay people (UK)	Lay people (Hamilton, Canada)	Lay people (UK)
Modelling					
Modelling approach	Statistical inference	Holistic	Statistical inference	Multi-attribute utility function	Statistical inference
Scoring algorithm	$1-(1-TTO)^{1.6}$	*Holistic*	1 + C1 + C2 + W1*Mobility + … + W5*Anxiety/depression	(1+C1)*(W1*Vision)*… *(W8*Pain) – C	1 + C1 + W1*Physical functioning + … +W6*Vitality

Subjective health estimation scale [94]. Disability and Distress Scale [117-118]. EQ-5D, EuroQol EQ-5D [103-105]. HUI3, Health Utilities Index version 3 [106]. SF-6D, Short Form Survey-6D [116]. W1 -8 are the weights for each item. C1 and C2 are constants.

The items for a health state classification system can be either designed explicitly for a utility-based instrument, or taken from a questionnaire for an existing profile-based instrument. The attributes for the Disability and Distress Scale, EQ-5D and HUI3 were designed explicitly with utility-based methods in mind. The number of items and their levels were deliberately restricted to limit the size and complexity of the instrument [2]. In contrast, the items for the SF-6D were taken from an existing 36-item generic profile-based instrument called the SF-36 [23]. The attributes for most disease-specific utility-based instruments are taken from existing disease-specific profile-based instruments [95, 97-98, 113].

Taking items from an existing instrument rather than designing a new instrument has a number of advantages. One is that the items often have established evidence of feasibility, reliability and validity. Another is that ratings already collected with the existing instrument in clinical trials and other studies can be converted to utility data, and future studies can provide both profile data and utility data with a single instrument to reduce respondent burden. A disadvantage of using an existing profile-based instrument is that it may contain large number of items that make the potential number of health states vert large, and the task of deriving the scoring algorithm very complex, as discussed above.

One way of limiting the complexity of a utility-based instrument that is derived from an existing profile-based instrument is to select a limited number of its items. For example, the SF-6D referred to above contains only six of the 36 items from the SF-36. The items can be selected by expert opinion, or psychometric methods that select items on the basis of criteria such as feasibility, discriminative ability, construct validity, and correlation with utilities [28, 97, 116]. Another way to reduce complexity is to use subscales derived from multiple items rather than individual items. The use of subscales may limit the ability of judges to understand individual health states, so the use of subscales is best restricted to situations where patients are judging the utility of their own health states and do not rely on the description of attributes to interpret it.

This section described methods to determine the items and response options of a utility-based instrument. The next section describes the valuation survey.

2.6.2 The valuation survey

The valuation survey involves valuing the health states that are described by the utility-based instrument, and takes place in three steps. This section adopts a classic description by Froberg and Kane [39]. The first step is to select an appropriate sample of 'judges' who will assign utilities to the health states. Typically this is a sample of the general population or patients. The second step is to present a set of health states to the judges, where each health state is described by its level of impairment on each attribute of the health classification system. The third step is for the judges to assign utilities to the health states presented to them using a direct utility-based scaling method such as the standard gamble or time trade-off. The three steps are now described in more detail.

Step 1 – Who to involve in the valuation task

The first step in performing the valuation survey is to select an appropriate sample of 'judges' who will assign utilities to the health states. As was discussed in section 2.5.4, the characteristics of the sample of judges who assign utilities to the health states in the valuation survey will determine the perspective of the utility-based instrument. Typically a sample of the general population or patients is used, but experts such as health-care professionals were used for early utility-based instruments [117]. It is important that a representative sample from the appropriate population is obtained. This is because the valuations depend on the experience, attitudes and beliefs of the judges [54]. Valuations often differ between patients and the general population, and even within each group because of differences in demographic background, social and cultural values, and political and economic systems; as was discussed in section 2.4.3. The sources of valuations for five commonly used utility-based instruments are shown in table 2.2. Three of the instruments source valuations from lay people in the United Kingdom or Canada, one from health care workers and one from patients.

Step 2 – Presentation of health states to judges

Once the sample of judges has been selected, the second step in the valuation survey is to present a set of health states to them. Each of the health states are described in terms of the level of impairment on each item of the instrument. The number of health states that are presented and their characteristics depends on the number of health states described by the instrument (table 2.2). For example, for a simple

instrument such as the Disability and Distress Scale, it is feasible to present all 29 health states. In contrast, for the SF-6D it is not feasible to present all 18000 health states. The approach used to produce the scoring algorithm (described in the next section) will determine which health states are presented. If the holistic approach is used to produce the scoring algorithm then all health states are valued. If statistical modelling is used, then only a limited number of health states are presented. The characteristics of the sample of judges perspective will also determine which health states are presented. If the general population is used, or patients are rating hypothetical health states, then the judges can assign a utility to any health state. If a sample of patients is used who are asked to assign a utility to their own health state, then those respondents can only consider that one health state. In this situation, it is important that the health states experienced by the sample of patients are diverse.

Step 3 – Scaling method

Once the set of health states have been presented to the judges, the third step in the valuation survey is for the judges to assign utilities to those health states. Respondents consider each health state that is presented, and assign a utility score between zero (death) and one (perfect health) using a direct utility-based scaling method. The standard gamble or time trade-off is typically used (table 2.2). The advantages and disadvantages of each scaling method were discussed previously in section 2.4.2.

The valuation survey results in utilities being assigned by one or more judges to a series of health states that are defined by levels of impairment on one or more items.

2.6.3 Modelling approaches to produce the scoring algorithm

The third stage is to produce a scoring algorithm that assigns a utility to every plausible health state described by the instrument. The three main modelling approaches to producing a scoring algorithm are described below.

The three modelling approaches for producing a scoring algorithm can be referred to as the ‘holistic’ approach, the ‘multi-attribute utility function’ approach, and the ‘statistical inference’ approach. The holistic approach requires all possible health states to be valued, and is only feasible for very simple health state classification systems. The other approaches require only a limited number of possible health states

to be valued, and are more feasible for more complex instruments that result in large numbers of potential health states. The latter approaches use statistical methods to examine the relationship between the utility of each health state and the level of impairment on each attribute. A statistical function is then derived that predicts the utility of any potential health state from all conceivable combinations of responses on each item [9, 39]. The modelling approaches for five utility-based instruments are shown in table 2.2, with the majority using the statistical inference approach. The process and merits of each approach are described in the next three sections.

Modelling by the holistic approach

The holistic approach was used in the earliest work on health utilities. It requires judges to value every health state derived by all conceivable combinations of different levels of each attribute. For example, the scoring algorithm for the Distress and Disability Scale, referred to previously, is called the Rosser Index and produced by establishing utilities for all 29 conceivable health states [117]. The advantage of the holistic approach is that it does not require any statistical modelling to produce a scoring algorithm. The major limitation of the holistic approach is the potential burden on judges if a large number of health states must be valued [39]. For example, establishing utilities for all 18000 health states described by the health state classification system of the SF-6D would be impractical. Statistical approaches have been developed to deal with valuations of large numbers of potential health states.

Modelling by the multi-attribute utility function approach

One approach that avoids having to value every health state is the multi-attribute utility function approach, also known as the explicitly decomposed approach. The evaluation process is broken up into a series of simpler subtasks. Subjects value health states with differing levels of impairment on each level of a single attribute, assuming the level of impairment on all other items are held constant. Few judgements where there is impairment on more than one item are required. These relate to corner states: for example, where one attribute is at its worst and all other attributes are at their best; and a limited number of multi-attribute states. The utility for a health state is estimated as a function of the underlying single items [39, 93, 106, 119].

The foundation of the multi-attribute utility function approach is multiattribute utility theory, which is an extension of von Neumann-Morgenstern utility theory that was discussed in section 2.4.1. Multiattribute utility theory states that there is no interaction between utilities among levels on any one attribute and the fixed levels for the other attributes. This assumption of multiattribute utility theory is called first-order utility independence. For example, worst impairment in physical function may have a utility of 0.6 on the physical function attribute regardless of the levels of impairment on the other attributes. Multiattribute utility theory allows utilities for each health state to be estimated as a function of the utilities of the underlying single attributes. The form of the mathematical function is dependent on the degree of independence between the attributes. It is often a multiplicative function, but may have a more complex additive or multi-linear function [39-40, 93, 106, 119-120]. Multi-attribute utility functions based on multiattribute utility theory have been produced for the Health Utilities Index (table 2.2) and Assessment of Quality of Life (AQoL) instruments [9, 106, 119, 121].

A strength of the multi-attribute utility function approach compared to the alternate statistical inference approach described in the next section is that it has a strong theoretical foundation in multi-attribute utility theory. Another strength over the statistical inference approach is that less health states need to be valued by judges in the valuation survey, because the focus of valuation is on a limited number of health states where there is impairment in only one attribute, and other attributes are not impaired [106, 119]. A potential limitation of the multi-attribute utility function approach compared with the statistical inference approach is that the type of health states presented to judges may not be credible. For example, a health state with one attribute at its worst and all other attributes at their best is unlikely to occur [116]. A more serious limitation is that the ability of a multi-attribute utility function to accurately predict utilities for health states may be inferior to that of the statistical inference approach discussed in the next section [119].

Modelling by the statistical inference approach

The third approach to produce a scoring algorithm is the statistical inference approach, also known as the econometric approach. Judgement is restricted to a limited number of health states, with differing levels of impairment on item. Utilities

of other health states are predicted using a statistical model. The model is derived by regression using data from the valuation survey. The utility is the dependent variable and the items are the independent variables. Models vary in which items from the instrument are included, and in whether the response levels for each item are combined. Each item can be represented as a continuous variable, or the shifts between response levels of an item can be represented by dummy variables [39, 116, 119]. The primary criterion for selecting one model over another is its ability to accurately predict a utility for a health state [116, 119].

The major advantage of the statistical inference approach over the multi-attribute utility function approach is that the resultant scoring algorithm may better predict utilities for health states, but a limitation is that there is no theoretical basis for the selected statistical model The selection of an appropriate statistical model is based purely on empirical findings, without reference to multi-attribute utility theory [119].

The statistical inference approach has been used to produce scoring algorithms for a number of multi-attribute instruments including the generic EuroQol EQ-5D [103] (table 2.2), Quality of Well Being scale [45], and Health Utilities Index-2 [119]; and several disease-specific instruments [97-98, 122]. The statistical inference approach can also be used to produce a scoring algorithm for an instrument derived from one or more single-item global scales [37, 91-92] (table 2.2).

2.7 Summary

This chapter has provided an overview of methods used to assess HRQL. It has focussed on the differences in measurement properties and applications of value-based and utility-based scaling methods, and the merits of instruments that focus on global HRQL or specific aspects. Methods to produce summary data from instruments were described that can be applied to evaluate treatments in clinical trials, and more broadly to inform decision-making about treatments. A utility-based instrument is one approach to producing summary data that combines aspects of both value-based and utility-based scaling methods, and uses both global and specific measures. A utility-based instrument is a practical way to determine the overall desirability of health states. To be valid and responsive, it is essential that a utility-based instrument asks about appropriate and relevant aspects of HRQL, and derives a utility index that provides an appropriate valuation of the desirability of a health state. Methods for developing a utility-based instrument to fulfil these requirements have been discussed.

The next chapter describes the study materials for the work developed in this thesis. The cancer-specific HRQL questionnaire is described from which a scoring algorithm will be derived that converts its responses to a utility index. The valuation survey used to produce the scoring algorithm, and clinical trials used to optimise, validate and apply the scoring algorithm are also described.

3. Study materials

3.1 Overview

This chapter describes the study materials and general methods used in this thesis. Section 3.2 describes a questionnaire about health-related quality of life called the Utility-Based Questionnaire-Cancer (UBQ-C), which is the focus of this thesis. The description outlines its purpose and conceptual basis, the history of its development, its composite items and their scoring, and its psychometric properties. Section 3.3 describes other questionnaires that were used to validate the indices derived from the UBQ-C. Section 3.4 describes the interview procedure used to directly elicit utilities from subjects. Sections 3.5 to 3.7 describe the designs, profiles, and patient characteristics of the three included studies used in this thesis. Sections 3.8 and 3.9 present the elicited ratings on the UBQ-C and other questionnaires, and utilities. The chapter concludes with a brief summary in section 3.10.

3.2 The Utility-Based Questionnaire-Cancer (UBQ-C)

This section describes the cancer-specific HRQL questionnaire from which a scoring algorithm is derived, optimised, applied and validated. The UBQ-C was completed by subjects in the valuation survey and trial datasets (section 3.5 below). As stated in section 1.1, the rationale for using this instrument was to facilitate the evaluation of a series of randomised controlled trials that have collected comprehensive HRQL data with this instrument. Content for this section is taken from published work by Martin et al [12-13].

3.2.1 Purpose

The UBQ-C is a disease-specific HRQL questionnaire that was designed to be an outcome measure for clinical trials in the field of cancer. Within this context, it was intended to serve two purposes. First, the UBQ-C was intended to be a cancer-specific profile-based questionnaire that can measure the effects of cancer and its treatment on a broad range of HRQL dimensions. This has been achieved in a range of cancer trials [123-127]. Second, the UBQ-C was intended to be a cancer-specific utility-based instrument that can derive utilities in clinical trials. This requires a scoring algorithm that converts the responses on the questionnaire to a utility index. The work presented in this thesis aims to derive such a scoring algorithm for the UBQ-C.

3.2.2 Conceptual basis

The UBQ-C was designed for use in clinical trials of cancer therapy, so it needed to be relevant to cancer patients, relatively brief, and easy to self-complete. It contains items that measure a cancer patient's experience of illness across all key dimensions of HRQL including general health, physical functioning, physical and psychological symptoms, role functioning, and social well-being. Additional items focus on the symptoms of cancer and the toxicity of treatment. Aspects of HRQL that are less likely to be influenced by cancer or its treatment such as deafness, blindness, or incontinence are not included; as is discussed in the next section.

3.2.3 History of development

The questionnaire was developed at the NHMRC Clinical Trials Centre, University of Sydney, Australia in parallel with a cardiovascular questionnaire called the Utility-Based Questionnaire-Heart (UBQ-H) [128-129]. The form and content of the

questionnaire builds on the conceptual framework for health status assessment developed by Gudex et al for an existing generic utility-based instrument called the Health Measurement Questionnaire [118].

The Health Measurement Questionnaire includes 36 items covering 5 key dimensions of HRQL (general mobility, usual activities, self-care activities, social and personal relationships, and psychological distresses). A generic core set of items from the Health Measurement Questionnaire was taken for the UBQ-C. Items less relevant for cancer patients (eg hearing, vision, writing, speaking and incontinence) were discarded, and the response formats of some items were modified.

Cancer-specific items were selected for addition by a review of existing literature on HRQL instruments used in cancer patients [130-131].

Two measures of global health status were also added. The health status thermometer is similar to the graduated, vertical, visual analog scale that accompanies the EuroQol EQ-5D questionnaire [103-105], but with the anchors of 'best imaginable health state' and 'worst imaginable health state' replaced by 'full health' and 'death', to conform with the requirements of a utility scale. The UBQ-C also includes the general health item from the Short-Form-36 health survey (SF-36) [132], which is a widely used and extensively validated measure of generic health status [5]. This item asks respondents to 'describe their general health' as excellent, very good, good, fair or poor.

3.2.4 Description

The UBQ-C includes 29 items about specific aspects of HRQL and two global scales (table 3.1). A sample of the questionnaire is reproduced in appendix 1.

The 29 items about specific aspects of HRQL are grouped into four subscales according to the HRQL dimensions that they are hypothesised to sample: physical function (3 items), social/usual activities (4 items), self-care (1 item), and distresses (21 items) due to physical and psychological symptoms associated with cancer and its treatment.

Table 3.1 Grouping of UBQ-C items within subscales

Subscale	Items
Physical function	Walking several blocks
	Climbing stairs
	Vigorous activities
Social/usual activities	Usual daily activities
	Social life
	Hobbies or leisure activities
	Sex-life
Self-care	Self-care
Distresses	Shortness of breath
	Difficulty sleeping
	Feeling sick (nausea/vomiting)
	Lack of energy
	Aches or pains
	Feeling sad or depressed
	Feeling anxious or worried
	Loss of appetite
	Dissatisfaction with your weight or appearance
	Uncertainty about the future
	Numbness or pins & needles
	Anger or resentment
	Loneliness
	Loss of hair
	Diarrhoea
	Constipation
	Loss of self confidence
	Feeling dependent on others
	Thought of chemotherapy
	Inability to concentrate
	Any other problems
Global scales	Health status thermometer
	General health

The items within the subscales for physical function, social/usual activities, and self-care each have four response categories: ‘not at all’, ‘slightly affected’, ‘severely affected’ and ‘unable to do activities at all’. The items within the distresses subscale have 11 response categories for amount of distress ranging from ‘0 (None)’ to ‘10 (Extreme)’ (appendix 1).

The two global scales are single items that ask respondents for a unified assessment of their health status and general health. The health status thermometer is a graduated, vertical, visual analog scale ranging from ‘100 (Full health)’ to ‘0 (death)’. The general health scale has five response categories: Excellent’, ‘Very good’, ‘Good’, ‘Fair’ and ‘Poor’.

3.2.5 Scoring

The scores for the subscales about physical function, social/usual activities, self-care and distresses are the simple average of the non-missing items, linearly transformed to a scale from 0 (worst) to 1 (best). More details are given in chapter 4 (section 4.5). Responses to the items labelled ‘Sex life’ and ‘Other problems’ are not included when calculating the scores for the subscales, because they are commonly omitted by respondents [12-13].

3.2.6 Psychometric properties

The psychometric properties of the UBQ-C in a cancer population have been reported previously. These include good feasibility (high completion rate with little missing data), internal consistency of subscales (Cronbach’s alphas > 0.75 and confirmatory factor analysis), test-retest reliability (intraclass correlation coefficients: median 0.85, lower quartile 0.81, upper quartile 0.90), convergent validity (substantial correlations with related instruments: GLQ-8 and GLQ-Uniscale [130], Priestman and Baum LASA scales [131], and Life Satisfaction Index-A [133]), discriminative ability (between groups with different disease severity) and responsiveness to change within individuals [12-13].

3.3 Other questionnaires about HRQL and health status

This section describes additional questionnaires that were completed by subjects and their clinicians. The ratings on these questionnaires are used to validate the utility index (chapters 5 and 6), and to inform treatment comparisons (chapters 6 and 7). A sample of each questionnaire is reproduced in appendix 2.

Three additional questionnaires were completed by subjects. The Spitzer-Uniscale of global life quality and the Priestman and Baum Linear Analog Self Assessment Scales (LASAS) were completed by subjects in the valuation survey and advanced cancer trial (section 3.5 below). The Chemotherapy Acceptability Questionnaire was completed by subjects in the advanced cancer trial only. The Spitzer-Uniscale is a validated, single-item visual analog scale that asks the respondent to indicate their 'overall life quality' [14-15]. The anchors of 'highest quality' and 'lowest quality' were replaced by 'best possible' and 'worst possible'. The LASAS is a validated measure of cancer-specific HRQL that is comprised of five visual analog scales comprised of horizontal lines about physical well-being, mood, pain, nausea and vomiting, and appetite [131, 134]. The Chemotherapy Acceptability Questionnaire is a study-specific scale for the advanced cancer trial that was designed to supplement the UBQ-C. It includes 15 items about the inconvenience and additional side effects that were expected to occur with the trial chemotherapy regimens but were not assessed by existing questionnaires.

Clinicians completed the Eastern Cooperative Oncology Group (ECOG) performance status scale in the advanced cancer trial. This validated scale rates patients' physical functional status as '0', fully active; '1', restricted in physical activity but able to do light work; '2', confined to a bed or chair for less than 50% of waking hours and capable of all self-care but unable to do any work; '3', confined to a bed or chair for more than 50% of waking hours but capable of limited self-care; and '4', totally confined to bed or chair, completely disabled, incapable of any self-care [135].

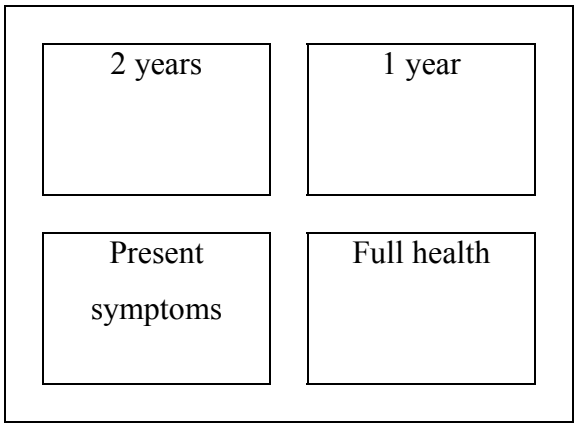
3.4 Interview procedure for utility elicitation

This section describes the interview procedure used to elicit utilities directly from subjects. Utilities were elicited in the valuation survey (section 3.5 below). The utilities are used to derive a scoring algorithm for the utility index (chapter 5).

Utilities were elicited directly from subjects about their current HRQL using the time trade-off technique with a hypothetical survival time of two years. The time trade-off technique was selected because it is practical, reliable, and has empirical validity (section 2.4.2). In particular, ease of administration is an important consideration for patients with advanced cancer. A two-year survival time was used because it is consistent with the expected median survival of the group. Face-to-face interviews were conducted by one trained researcher and took about 30 minutes to complete. A script was used to standardise the administration of each interview, and a series of cards were used as visual aids to accompany each question. The subjects were presented with pairs of hypothetical situations and asked to indicate which was preferable (figure 3.1). The time spent in full health was changed (ping-ponged) until a point of indifference was reached. Subjects were reassured that the hypothetical choices were not meant to reflect their present circumstances and that their answers would have no effect on their future medical care.

Figure 3.1 Time trade-off interview: script and visual aid

"The situation on the left involves living for 2 years with symptoms identical to the ones you are presently experiencing. The symptoms are stable - they don't get any worse or any better. The other situation involves living less time, 1 year in full health. Think about these two situations and tell me which one you feel is most [sic] preferable"



3.5 Included studies

This section describes the three studies that provide data for the work presented in this thesis. The 'valuation survey' was a cross-sectional study of patients with advanced cancer. Participants completed the UBQ-C and other questionnaires, and assigned utilities in a face-to-face interview. This data is used to derive the scoring algorithm (chapters 4 and 5). The 'advanced cancer trial' and 'early cancer trial' were two randomised clinical trials of chemotherapy for breast cancer. Participants completed the UBQ-C and other questionnaires at various time points during the trials. This data is used to optimise, validate and apply the scoring algorithm (chapters 6 and 7). Details of each study are given in the following section.

3.5.1 Valuation survey

The valuation survey was a cross-sectional study of ambulatory patients who were recruited from two tertiary-referral oncology outpatient units. Eligible patients had advanced cancer, impaired HRQL, and were willing and able to complete both a self-administered HRQL questionnaire and a one hour interview in English [13].

Potential subjects were given a patient information sheet (appendix 3). Consenting subjects were registered and scheduled for an interview, usually on the day of their next appointment at the oncology clinic. Utilities were elicited directly from subjects about their current HRQL by one trained researcher using a standardised, face-to-face, time trade-off interview with a hypothetical survival time of two years (section 3.4). The time trade-off was expressed on the standard continuum where 1 represents full health and 0 represents dead [16]. Patients were mailed the UBQ-C (section 3.2 and appendix 1), Spitzer-Uniscale and LASAS (section 3.3 and appendix 2), and asked to complete them 3 to 7 days before the planned interview [13].

3.5.2 Advanced cancer trial

The advanced cancer trial was conducted by the Australian New Zealand Breast Cancer Trials Group and included patients with advanced breast cancer who were randomly allocated to receive either daily oral capecitabine or standard cyclophosphamide, methotrexate and 5-fluouracil (CMF) as first-line chemotherapy until disease progression. The primary outcome measure of the trial was quality-adjusted time to progression. Secondary outcome measures were time to progression,

response rates, HRQL, overall survival, safety and cost-effectiveness. Eligible subjects were 18 years or older, and were about to start first-line chemotherapy for histologically confirmed advanced breast cancer. Subjects were excluded if they were totally confined to bed and completely disabled (ECOG performance status 4, as described in section 3.3). Enrolment was from June 2001 to July 2005 at 34 centres in Australia and New Zealand [124].

Potential subjects were given a patient information sheet (appendix 4). Consenting subjects completed the UBQ-C (section 3.2 and appendix 1), and the Spitzer-Uniscale, LASAS and Chemotherapy Acceptability Questionnaire (section 3.3 and appendix 2) unless they could not read English. Clinicians completed the ECOG performance status scale (section 3.3 and appendix 2). Questionnaire were completed at baseline (prior to randomisation), then every three to four weeks during treatment and until disease progression. Questionnaires were not completed after disease progression. For the analyses reported in chapter 7, 'during treatment' was defined as from the time of randomisation until 30 days after disease progression.

The data from baseline questionnaires competed prior to randomisation was used to optimise, apply and validate the scoring algorithm (chapter 6), and the data during treatment was used to apply the scoring algorithm to a treatment comparison (chapter 7).

3.5.3 Early cancer trial

The early cancer trial was conducted by the Australian New Zealand Breast Cancer Trials Group in collaboration with the International Breast Cancer Study Group. It included patients with high-risk early stage breast cancer who were randomly allocated to receive either high-dose chemotherapy with stem cell support over 12 weeks or standard-dose chemotherapy over 24 weeks. The primary outcome measure of the trial was overall survival. Secondary outcome measures were quality-adjusted survival, disease-free survival, toxicity, HRQL and cost-effectiveness. Eligible subjects were aged 16 to 65 years, and were about to start adjuvant chemotherapy for histologically confirmed early-stage primary breast cancer with 5 or more involved axillary lymph nodes. Subjects were excluded if they were capable of only limited self-care and/or were confined to a bed or chair for more than 50% of waking hours

(ECOG performance status 3 or 4). Enrolment was from March 1997 until March 2000 at multiple centres in Australia, New Zealand, Europe and Asia [136].

Potential subjects were given a patient information sheet (appendix 5). Consenting subjects living in Australia and New Zealand were eligible to participate in a substudy about HRQL and resource usage. Substudy participants were required to complete the UBQ-C prior to starting chemotherapy (baseline), 12 weeks after randomisation (during chemotherapy), and a few months after completing chemotherapy.

3.6 Study profiles

This section describes the study profiles for the three studies. For the valuation survey, 204 of the 323 patients that were approached to take part in the study were eligible (figure 3.2). Compliance was excellent, with planned interviews and questionnaires completed by 98% of participants. For the advanced cancer trial, compliance was excellent with questionnaires completed by over 95% of subjects who were expected to complete them (figure 3.3). For the early cancer trial, compliance was not as good with questionnaires completed by 72% prior to chemotherapy, 40% during chemotherapy, and 88% after completing it (figure 3.4). Good compliance is important in HRQL assessment, because patients who do not comply with assessment tend to have worse HRQL [137]. The resultant missing data can bias results of HRQL analyses in favour of patients with better HRQL.

Figure 3.2 Valuation survey: study profile

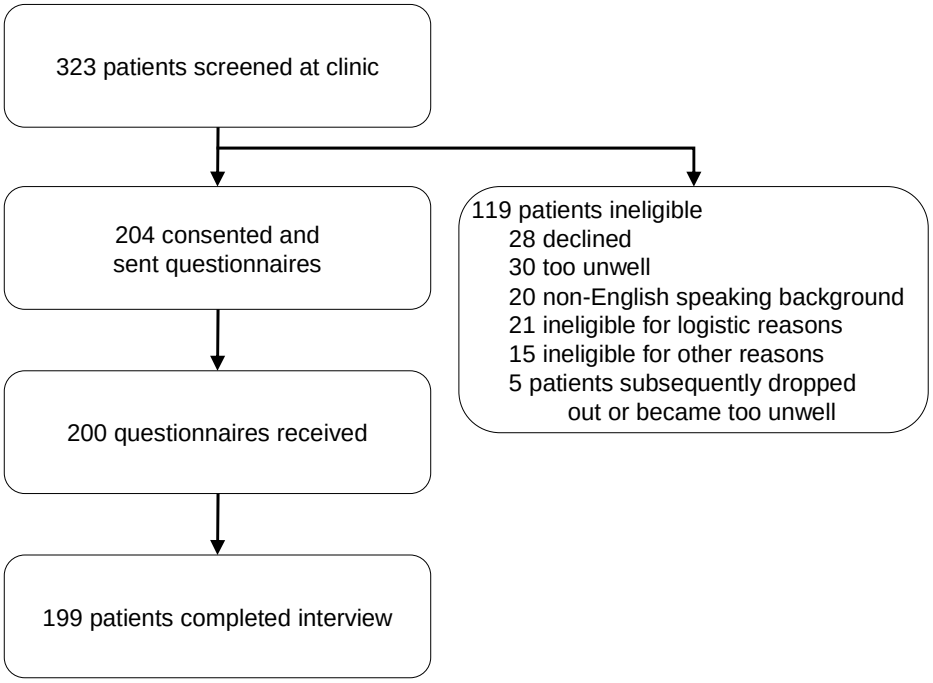


Figure 3.3 Advanced cancer trial: study profile

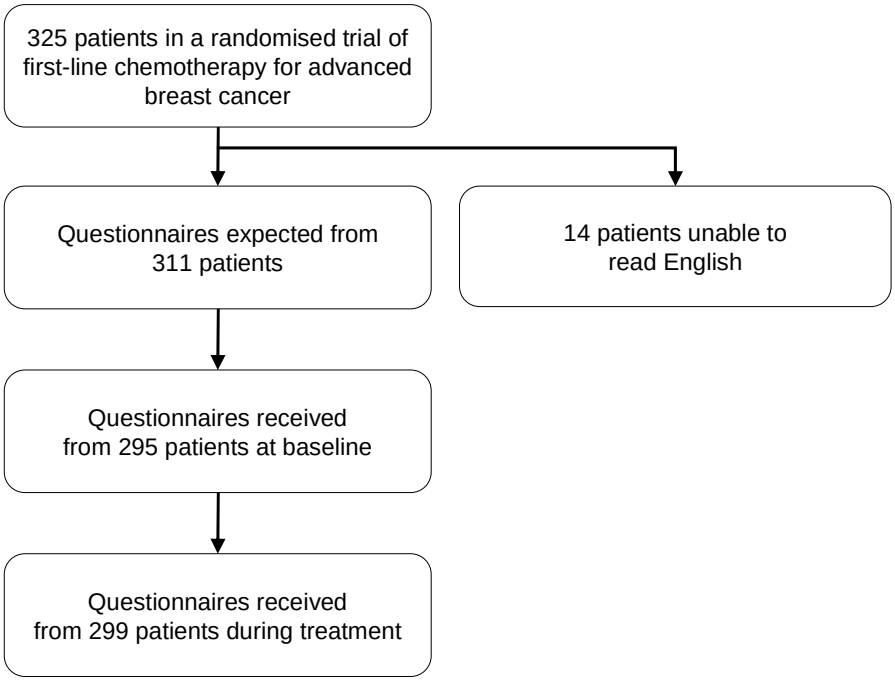
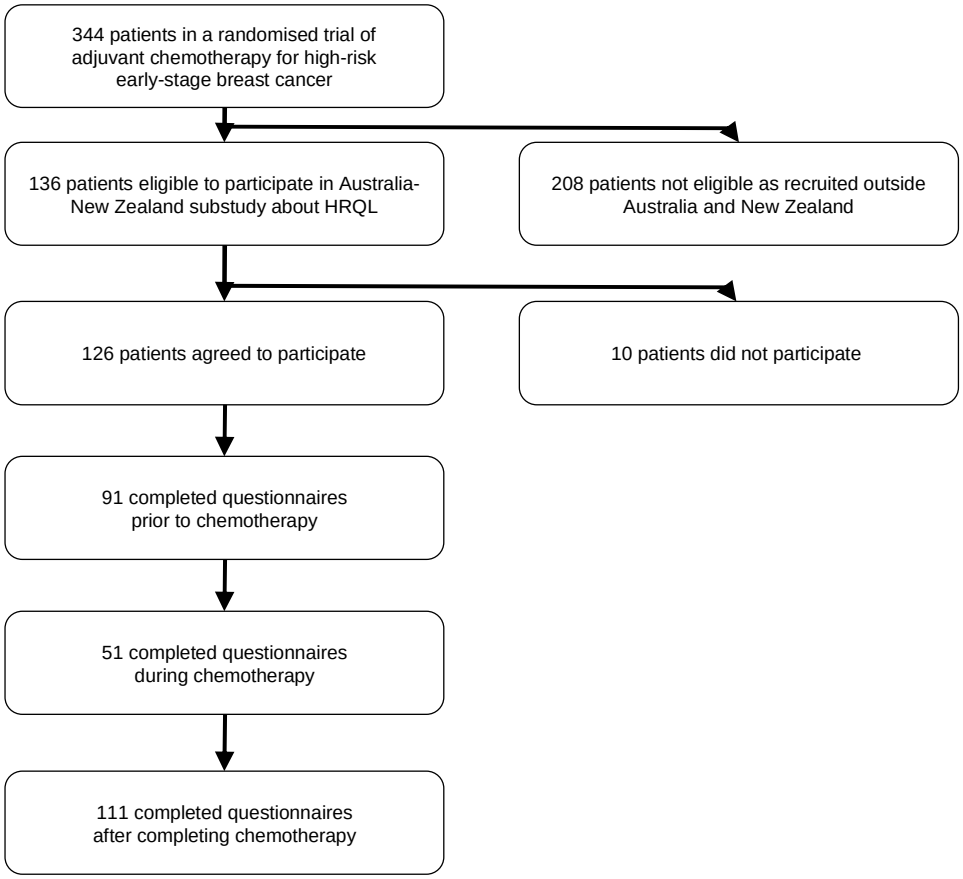


Figure 3.4 Early cancer trial: study profile



3.7 Patient characteristics

This section describes the characteristics of the 655 subjects in each of the three studies (see table 3.2).

The valuation survey consisted of 204 patients with advanced cancer, mostly arising from the breast or bowel. The other cancer types were lung (2), ovarian (5), sarcoma (5), melanoma (4), prostate (3), and non-Hodgkins lymphoma (3). Male and female patients were included and most age groups were represented (range 22 to 81 years, mean 56 years). The different levels of self-reported general health status were well-represented. Most modes of treatment were represented: chemotherapy, supportive care and observation.

The trial datasets consisted of 451 patients with breast cancer of both early and advanced stages. All subjects were female and most age groups were represented (range 25 to 84 years). For the advanced cancer trial, most had good performance status (ECOG 0 in 34% and ECOG 1 in 54%), and fewer had poor performance status (ECOG 2 in 11% and ECOG 3 in 2%). The different levels of self-reported general health status were well-represented.

Table 3.2 Included studies: patient characteristics

Dataset	Valuation survey (n=204)	Advanced cancer trial (n=325)	Early cancer trial (n=126)
Cancer stage	Advanced	Advanced	High-risk early-stage
Cancer type (%)			
Breast	50	100	100
Bowel	29	-	-
Other	21	-	-
Gender (%)			
Male	32	-	-
Female	68	100	100
Age (Years) (%)			
< 40	12	2	14
40-49	21	12	47
50-59	25	29	35
60-69	28	36	3
≥ 70	14	21	-
General health at baseline (%)			
Excellent	10	6	22
Very good*	-	18	-
Good	48	30	54
Fair	35	32	19
Poor	7	13	4

* Response category 'Very good' not included in some versions of 'General health' item of UBQ-C

3.8 Ratings on the UBQ-C

This section describes and compares the ratings on the UBQ-C for each of the three included studies.

All items on the UBQ-C except 'Sex life' and 'Other problems' were completed by over 90% of patients in all studies. Ratings on the health status thermometer and subscales are summarised in table 3.3. Patients in each study consistently reported worst impairment for physical function and least impairment for self-care.

The ratings from the two trial datasets were compared. At baseline, patients with advanced cancer reported worse health status than patients with early cancer, as expected (means of 0.69 vs 0.81, difference 0.13 [with rounding], 95% CI 0.08 to 0.17, $p<0.0001$). Patients with early cancer reported worse health status during chemotherapy than before starting it (means 0.68 vs 0.81, mean deterioration 0.13, 95% CI 0.08 to 0.19, $p<0.0001$); or after finishing it (means 0.68 vs 0.84, mean improvement 0.15 [with rounding], 95% CI 0.10 to 0.21, $p<0.0001$). Similar differences were reported for ratings on UBQ-C subscales (table 3.3). Ratings during treatment for the advanced cancer trial are not reported here, but are reported in chapter 7.

Table 3.3 Included studies: ratings on UBQ-C

Dataset	Valuation survey		Advanced cancer trial		Early cancer trial					
Cancer stage	Advanced		Advanced		High-risk early-stage					
Treatment phase	Various		Before treatment		Before treatment		During treatment		After treatment	
n	204		295		91		51		111	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Health status thermometer	0.74	0.16	0.69	0.20	0.81	0.15	0.68	0.21	0.84	0.13
UBQ-C subscales										
Physical function	0.65	0.23	0.53	0.32	0.77	0.21	0.63	0.24	0.80	0.20
Social/usual activities	0.77	0.22	0.66	0.29	0.74	0.23	0.69	0.22	0.88	0.17
Self-care	0.95	0.14	0.89	0.20	0.89	0.15	0.97	0.10	0.99	0.06
Distresses	0.80	0.15	0.78	0.15	0.77	0.15	0.69	0.18	0.83	0.13

SD, standard deviation. UBQ-C, Utility-Based Questionnaire-Cancer. HRQL, health-related quality of life.

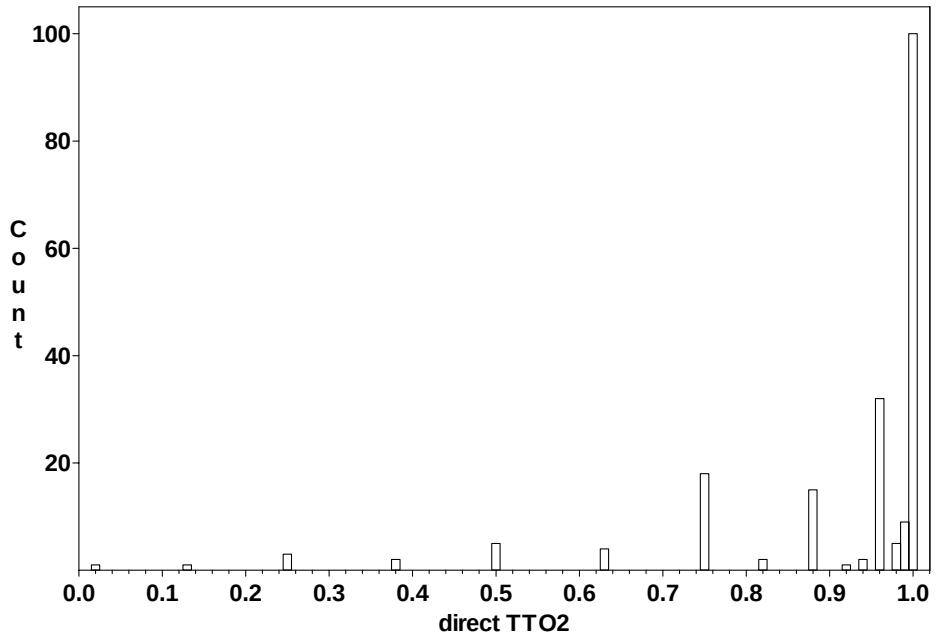
All ratings on scale from best (one) to worst (zero).

3.9 Elicited time trade-off utilities

This section describes the time trade-off utilities that were assigned by subjects in the valuation survey.

The mean (SD) time trade-off utility assigned by subjects in the valuation survey about their own health state was 0.91 (0.17), and the 95% confidence intervals were 0.89 to 0.94. The median of the time trade-off utilities was 0.995. The time trade-off utility was tied at 1.000 for 100 of 200 subjects, despite their significant impairments in HRQL as self-reported on the UBQ-C. Because of the skewed distribution and the spike at 1.000, the mean value of the time trade-off utility was lower than its median value (figure 3.5). The implications of skewed and spiked data are discussed in chapter 4.

Figure 3.5 Valuation survey: histogram of time trade-off utilities



directTTO2, 2-year time trade-off utility

3.10 Summary

This chapter described the study materials used in this thesis. The comprehensiveness, conceptual basis and psychometric properties of the Utility-Based Questionnaire-Cancer (UBQ-C) were presented along with a description of other questionnaires used in this thesis. The three included studies used in this thesis were described in detail, together with the characteristics of the participants and their ratings on the UBQ-C, other questionnaires, and time trade-off utilities. The next chapter outlines the general approach used for the work presented in this thesis.

4. Statistical methods

4.1 Overview

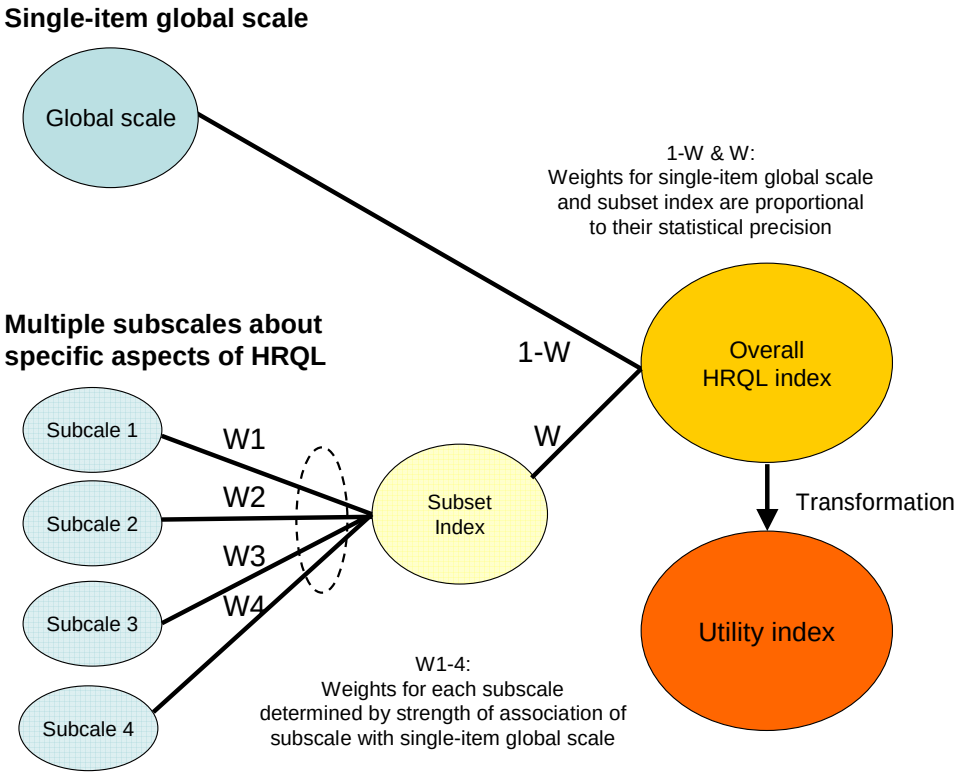
This chapter describes the general methods used in this thesis to develop a scoring algorithm that converts the responses to the UBQ-C cancer-specific questionnaire (described in section 3.2 of chapter 3), into an optimally-weighted utility index. Section 4.2 provides the rationale for the methodological approach taken. Section 4.3 describes the general approach. This involved deriving, optimising, validating and applying the scoring algorithm. Sections 4.4 to 4.9 describe the methods for statistical analysis. The chapter concludes with a brief summary.

4.2 Background

The methodological approach taken in this thesis is novel in that the utility index was derived from both a single-item global scale, and multiple subscales about specific aspects of HRQL. In contrast, most utility indices are derived from either multiple subscales, or a single-item global scale, but not both. The conceptual and empirical arguments supporting each approach were described in chapter 2 (sections 2.3.2 and 2.5.2). The rationale for deriving a utility index from both multiple subscales and a single-item global scale is to incorporate the strengths of each approach, and is advocated by Lumley et al [32]. The single-item global scale provides a unified assessment of global HRQL. It may capture additional aspects contributing to global HRQL that are important but not detected by specific subscales. The inclusion of multiple subscales may improve the precision of the index, because it is derived from multiple items, and explicitly assesses specific aspects of HRQL. Lumley presented evidence that the resultant index has better reliability than a single-item global scale without meaningfully altering validity [32]. Lumley has also developed a novel method to produce the scoring algorithm for such an index, which is described below.

The scoring algorithm was produced using what others have referred to as the statistical inference approach, whereby valuation is restricted to a limited number of health states with differing levels of impairment on each scale, and utility scores for other health states are predicted using a statistical model. This approach was the most feasible for the data obtained from the valuation survey, because of the diversity of health states that were represented. The strengths and weaknesses of alternatives were outlined in chapter 2 (section 2.6.3). Lumley's approach is a novel variant of the statistical inference approach that is designed to optimally combine a single-item global scale and multiple subscales [32]. The three steps of Lumley's approach are outlined in figure 4.1, and described in the following paragraphs.

Figure 4.1 Deriving a utility index with Lumley's combined approach



The first step of Lumley's approach calculates a 'subset index' (referred to by Lumley as the subset estimate) by combining multiple scales about specific aspects of HRQL according to weights based on their correlations with the single-item global scale. The weights are proportional to the coefficients from a multivariable linear regression of the single-item global scale on the specific subscales.

The second step calculates an 'overall HRQL index' (referred to by Lumley as the global estimate) by combining the subset index with the single-item global scale. Extra weight is given to the component with less measurement error. The measurement error of the subset index is taken to be the error mean square of the linear regression of the global scale on the subscales. The measurement error of the global scale is taken to be the variance of the global scale multiplied by (1- its intra-class correlation coefficient). The intra-class correlation coefficient is a measure of test-retest reliability. It is a generalisation of Pearson's correlation that measures absolute agreement amongst two or more ratings [138]. A more reliable scale has a higher intra-class correlation and less measurement error. The weight for the subset index is derived by dividing the measurement error of the subset index by the measurement error of the global scale. The weight for the global scale is 1 – the weight for the subset index.

The third step calculates a utility index from the overall HRQL index that is expressed with a utility-based scaling method. Lumley recommends a transformation function of the form used to convert single-item global scales such as a visual analogue scale to a utility index [32]. Potential models that include power transformations, linear, quadratic and cubic models [37] are discussed in more detail in the next chapter. The transformation function is derived in a valuation survey, as described in chapter 2 (section 2.6.2). Respondents rate health states by assigning utilities with a direct utility-based scaling method (eg. time trade-off), and by assigning responses to the single-item global scale and the multiple subscales.

One novel aspect of Lumley's approach described above is in its combination of a single-item global scale with multiple subscales for specific aspects of HRQL. Another is that the scoring algorithm can be optimised in different clinical contexts by adjusting the weights for the multi-item subscales. The purpose of optimising the

algorithm is to reflect the differences in importance that patients with different types and stages of disease and treatment assign to various dimensions of HRQL [32, 63-64, 139]. In contrast, standard utility-based instruments rely on fixed weights that are derived from populations that may be very different [64].

The next section describes how Lumley's approach was adapted to develop a scoring algorithm for the UBQ-C.

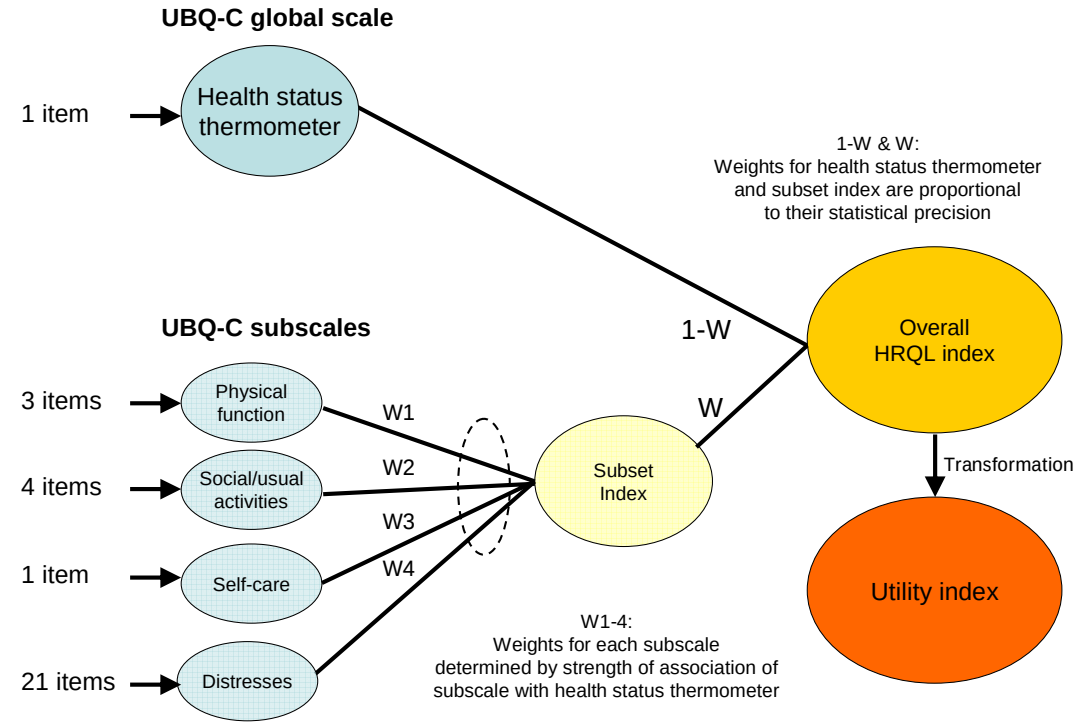
4.3 General approach

This section describes the general approach taken in this thesis to develop the UBQ-C as a cancer-specific utility-based instrument. This involved deriving, optimising, validating and applying a scoring algorithm that converts ratings from the UBQ-C to indices of overall HRQL and utility.

The scoring algorithm was derived by adapting Lumley's approach and applying it to the UBQ-C. The first step determined the scales from which the utility index would be derived. This is analogous to developing a health state classification system for a utility-based instrument, as described in chapter 2 (section 2.6). The health status thermometer was the single-item global scale. The multi-item subscales for physical function, social/usual activities, self-care and distresses were the multiple subscales. The general health item of the UBQ-C was not included, because it is strongly correlated with the health status thermometer and is unlikely to contribute additional information [12]. The four multi-item subscales were used rather than the individual items of the UBQ-C to minimise problems of collinearity.

The second step calculated a subset index, overall HRQL index and utility index from a weighted combination of the health status thermometer and multi-item subscales. This process is analogous to valuing a subset of the health states described by a health state classification system, as described in chapter 2 (section 2.6). Data was used from the valuation survey of patients with advanced cancer referred to in chapter 3 (section 3.5.1). The weights were derived by applying Lumley's approach, as above. The method is illustrated in figure 4.2, and detail of the methods and results are presented in chapter 5.

Figure 4.2 Deriving a utility index for the UBQ-C with Lumley's combined approach



UBQ-C, Utility-Based Questionnaire-Cancer. HRQL, health-related quality of life.

The third step optimised the scoring algorithm for two clinical contexts by adjusting the weights assigned for the health status thermometer and the multi-item subscales. The contexts were adjuvant chemotherapy for early breast cancer, and palliative chemotherapy for advanced breast cancer. Data from the early trial and advanced trial (referred to in sections 3.5.2 and 3.5.3 of chapter 3) were used to adjust the weights. The utility index was then validated by applying the scoring algorithm to the questionnaire data from these trials, and comparing the scores with responses to other questionnaires and patient characteristics. Details of the methods and results are presented in chapter 6.

The fourth step applied the scoring algorithm to evaluate the differences in utility between chemotherapy regimens in these clinical contexts. The regimens were high-dose versus standard adjuvant chemotherapy for early breast cancer, and single-agent oral versus multi-agent intravenous chemotherapy for advanced cancer. The analyses used cross-sectional data from the early breast cancer trial (during the intense phase of treatment), and longitudinal data from the advanced breast cancer trial (during the entire period from randomisation to disease progression). The latter required the scoring algorithm to be adapted to the analysis of longitudinal data, and extended by incorporating additional information about HRQL. Detail of the methods and results are presented in chapters 6 and 7.

4.4 Missing and censored data

There was a limited amount of missing data within the datasets analysed in this thesis. Some questionnaires were not submitted, some interviews were not completed, and some items within submitted questionnaires were left blank. Details were reported in the study profiles in chapter 3 (section 3.6 and 3.8, figure 3.2).

Missing data within questionnaires was considered missing at random. The scores on the UBQ-C subscales were imputed as the average of the non-missing items, when some (but not all) items from a subscale were missing. Analyses involving multiple subscales and the health status thermometer only included individuals with complete data. Individuals without complete data were excluded.

Some of the subjects in the advanced cancer trial had missing HRQL data prior to progression. These were assumed to be missing at random. No specific adjustments were made to account for these missing values. In effect, this means that the missing values were assumed to be similar to the average of the same patient's non-missing values.

4.5 Data transformations

Responses to the questionnaires were transformed for the analyses reported in chapters 5 to 7. The methods for each questionnaire are described below.

UBQ-C

For the UBQ-C, the responses to the items were converted to an interval scale ranging from 0 (worst) to 1 (best) as follows:

The responses to the items within the subscales for physical function, social/usual activities and self-care were converted from four ordinal response levels of 'Not affected', 'Slightly affected', 'Severely affected' and 'Unable to do activities at all' to 0, $^1/_3$, $^2/_3$, and 1 respectively. The responses to the items within the distresses subscale were converted from an 11-point scale (0 to 10) by dividing by 10. The score for each of the four multi-item subscales was the simple average of the non-missing component items, but the responses to the items labelled 'sex life' and 'other problems' were excluded because they are commonly left blank by respondents (table 3.1, section 3.2.5 of chapter 3).

The responses to the health status thermometer were converted from an interval scale ranging from 0 (worst) to 100 (best), by dividing by 100.

The format of the general health item differed in each study. For the valuation survey, the responses to the general health item were converted from four ordinal response levels of 'Excellent, 'Good', 'Fair' and 'Poor' to 0, $^1/_3$, $^2/_3$, and 1 respectively. For the early and advanced cancer trials, the responses were converted from five ordinal response levels of 'Excellent, 'Very good', 'Good', 'Fair' and 'Poor' to 0, $^1/_5$, $^2/_5$, $^3/_5$, $^4/_5$ and 1 respectively.

The transformed health status thermometer, general health item, and 4 multi-item subscales were assumed to have interval properties. This is common practice, but was not empirically tested

Spitzer-Uniscale, and Priestman and Baum Linear Analog Self Assessment Scales

The responses to the Spitzer-Uniscale and the Priestman and Baum Linear Analog Self Assessment Scales (LASAS) were converted from an interval scale ranging between 0 (best) and 100 (worst) to a scale from 0 (worst) to 1 (best) by subtracting each score from 100, and dividing the result by 100.

Performance status scale

The responses to the Eastern Cooperative Oncology Group (ECOG) performance status scales were converted from five ordinal response levels of '0' to '4' to dichotomous response levels of 'good' for response levels 0 and 1, and 'poor' for response levels 2, 3 and 4.

Time trade-off interview

The responses to the time trade-off interview were converted from a number representing the number of years in full health that was equivalent to two years with present symptoms, to a utility value ranging between '0' (death) and '1' (full health), by dividing each response by two.

4.6 Measures of central tendency

The responses to the utility interviews, and to a lesser extent the questionnaires, suffered from a ceiling effect with many responses at the upper end of the scale (full health), as reported in chapter 3 (sections 3.8 and 3.9). 50% of the respondents to the utility interviews assigned a utility of 1.0. The effect of this skewed distribution was that the median was higher than the mean. This raised questions about whether to use the median or mean as the measure of central tendency for reporting on groups, and for regression analyses.

Statistical arguments favour use of the median, but the mean is generally recommended for applications based on economic methods [116]. The statistical argument for using the median for a skewed distribution is that the mean is unduly affected by the outlying observations [138]. The philosophical argument for using the mean is that all responses are 'acceptable' and should contribute equally including extreme values, whereas using the median filters outlying values which are implied to be 'unacceptable' [140]. This argument is based on the principles of utilitarianism [140] and is founded in welfare economic theory [141]. For this reason, health economists generally recommend that applications based on economic methods such as utilities should be aggregated using the mean, irrespective of the skewness of that distribution [141]. This was the approach used in this thesis, but arguments advocating use of the median are acknowledged [141-143].

4.7 Statistical tests

Parametric tests were used to assess the statistical significance of differences in mean scores, despite the lack of normality of distributions for responses to the utility interviews and questionnaires. This was done in accordance with the recommendations of economists to use the mean in preference to the median, as described in the previous paragraph. The alternative would have been to compare distributions using non-parametric tests, but the results and conclusions were similar.

The statistical significance of differences in scores between groups were assessed with unpaired t-tests [138]. The statistical significance of differences in scores within groups from one timepoint to another were compared with paired t-tests [138]. The statistical significance of differences in scores within groups over multiple timepoints were calculated with generalised estimating equations that took into account the correlation between successive observations for each subject. For these analyses, the 'PROC GENMOD' statement of SAS was invoked, and a normal distribution was specified with the identity link function. Reported p-values for all statistical tests were two-sided.

The strengths of associations between related measures on scales were calculated using Spearman's rank-order correlation [138].

The relative precision of related measures was compared with the relative efficiency statistic [138, 144]. The reciprocal of the relative efficiency statistic is the factor by which the sample size can be reduced when a more precise and therefore more efficient scale is used. The relative efficiency statistic was calculated as the squared ratio of the t-score from the comparison of the groups using the measure under evaluation divided by the t-score from the comparison of the groups using the reference measure.

The statistical significance of differences in survivals between groups were compared with the logrank test, and the survival curves for each group were derived using the Kaplan-Meier method [138].

4.8 Regression models

Standard linear regression models based on minimising the differences between the observed and predicted values using ordinary least-squares regression were used in chapters 5 to 7 to derive the scoring algorithm. First, multivariable linear regression was used to calculate scores on the subset index from a weighted combination of the UBQ-C subscales according to their correlations with scores on the health status thermometer. Second, information from that regression was used to combine the subset index with the health status thermometer. Third, a disutility power transformation was selected amongst other candidates to map the relationship between the overall HRQL index and the time trade-off utility interviews. Fourth, this power transformation was used to convert the overall HRQL index to the utility index (figure 4.2).

4.9 Statistical software

All analyses were performed using SAS for Windows Release 8.02 [145].

4.10 Summary

This chapter has described the statistical methods used to develop a utility index for the UBQ-C. The background described the method, rationale and merits of Lumley's approach. This approach combines a global scale and multiple specific subscales, and optimises the combination in specific clinical contexts. Next, the general approach taken to the work reported in this thesis was described. This involved deriving, optimising and validating the scoring algorithm for the utility index. The remainder of the chapter outlined the methods for statistical analysis.

The next chapter reports the results of the approach taken to derive the scoring algorithm for the utility index.

5. Deriving a patient-based cancer utility index from a cancer-specific quality of life questionnaire

5.1 Overview

This chapter is a published work. The entire manuscript is quoted verbatim, and amendments are presented in italics. The supplementary online appendix for the manuscript is quoted verbatim in the supplementary section to this chapter (section 5.6). This supplementary section describes additional analyses that support the work reported in the manuscript.

Publication details

Grimison PS, Simes RJ, Hudson HM, Stockler MR.

Deriving a patient-based utility index from a cancer-specific quality of life questionnaire. Value in Health 2009; 12(5):800-807 (PMID 19508665).

Contribution of authors

PSG developed the research proposal, selected the research methods, did data analysis, interpreted the findings, and drafted the manuscript.

RJS conceived the research proposal, participated in selection of research methods and interpretation of findings, and contributed to the drafting and revision of the manuscript.

HMH provided guidance for data analysis and interpretation of the findings, and contributed to the drafting and revision of the manuscript.

MRS contributed to the conception and development of the research proposal, selection of research methods, interpretation of findings, and drafting and revision of the manuscript.

Abstract

Objectives: The aim of this study was to derive a scoring algorithm for a validated disease-specific quality of life instrument called the Utility-Based Questionnaire-Cancer (UBQ-C) that provided a utility index designed to inform clinical decisions about cancer treatments.

Methods: The UBQ-C includes a scale for global health status (1 item); and subscales for physical function (3 items), social/usual activities (4 items), self-care (1 item), and distresses due to physical and psychological symptoms (21 items). A scoring algorithm was derived to convert the subscales into a subset index, and combine it with the global scale into an overall HRQL index, which was converted to a utility index with a power transformation. The valuation survey consisted of 204 advanced cancer patients who completed the UBQ-C and assigned time-trade-off (TTO) utilities about their own health state. Preliminary validation involved comparing these derived utilities with other measures of HRQL.

Results: Weights for the subset index were: physical function 0.28, social/usual activities 0.06, self-care 0.01, and distresses 0.64. Weights for the overall HRQL index were health status 0.65 and subset index 0.35. The mean of the utility index scores was similar to the mean of the TTO utilities (0.92 vs 0.91, p=0.6). The utility index was substantially correlated with other measures of HRQL.

Conclusions: Data from a simple, self-rated, disease-specific questionnaire can be converted into a utility index suitable for comparing the net effect of cancer treatments on quality of life, and to evaluate trade-offs between quality and quantity of life in quality-adjusted survival analyses.

5.2 Introduction

Utility-based instruments are a common means of generating utility scores for calculating quality-adjusted life-years (QALYs) [9]. A utility-based instrument generally consists of a questionnaire which elicits responses about multiple dimensions of health status and health-related quality of life (HRQL), and a scoring algorithm that is used to convert the ratings on the questionnaire into a single utility-based index [9, 90]. The scoring algorithms for utility-based instruments are valued in surveys, where subjects are asked to assign utilities to the health states defined by the questionnaire [9, 90]. For example, the valuation survey for the EQ-5D instrument involved lay people assigning utilities with the time trade-off method to a number of hypothetical health states defined in five generic dimensions (mobility, self-care, usual activities, pain/discomfort and anxiety/depression)[100, 103]. Utility-based instruments may vary in the type of questions they contain (generic or disease-specific), and the perspective from which the scoring algorithm is valued (lay people or patient) [9, 36, 97].

Controversy exists about the suitability of generic versus disease-specific utility-based instruments for generating utility scores. Generic instruments like the EuroQol EQ-5D [103], Health Utilities Index (HUI3) [106] or SF-6D [107] ask about core aspects of HRQL that are of interest in a wide range of settings. The main argument for using a generic utility-based instrument is that it allows comparisons across a wide range of diseased and healthy populations [68, 108-109]. However a generic instrument is likely to provide an inadequate description of many diseases, so the utility scores that it generates may be insensitive to differences between individuals with that disease [108, 110-112]. Disease-specific, utility-based instruments were designed to address this lack of sensitivity by asking about specific aspects of HRQL relevant to that disease or condition [97-98, 110, 113].

Controversy also exists about the suitability of utilities that are valued from the perspective of lay people versus patients [58-59, 68]. The distinction is important because a patient typically assigns a higher utility to a health state than a lay person [56-58]. Economic guidelines generally recommend the use of generic utility-based instruments based on the perspective of lay people [69-71]. The main argument for

using the perspective of lay people for informing funding and policy decisions is that the primary objective in a publicly funded health system is to maximise health for society [9]. It is generally recommended that the perspective of patients is used to inform clinical decision making [9, 16, 36, 61]. The main argument for using the perspective of patients for clinical decisions is that the primary objective is to maximise health for the individual patient experiencing that condition [62, 73].

We posit that disease-specific instruments valued by patients are preferable for informing clinical decisions, whereas generic instruments valued by lay people may be preferable for decisions about health policy.

The aim of this study was to derive a scoring algorithm for a disease-specific, utility-based, HRQL instrument that is designed to inform clinical decisions about cancer treatments. The algorithm converts ratings from a cancer-specific HRQL questionnaire into a utility-based index designed to reflect the perspective of cancer patients. This paper describes the development and preliminary validation of the algorithm. A companion paper describes the application of the algorithm to trial datasets, and illustrates how it can be optimised in different treatment contexts [146] (*Chapter 6*).

5.3 Methods

5.3.1 Source of data

The valuation survey used to derive the scoring algorithm involved ambulatory patients with advanced cancer and impaired HRQL who were recruited from two tertiary-referral oncology outpatient units (*as described in section 3.5.1 of chapter 3*) [13]. Eligible patients had advanced cancer, impaired HRQL and were willing and able to complete a self-administered HRQL questionnaire and participate in a one hour interview in English. All patients provided written informed consent. The study was approved by the human research ethics committees at all participating institutions.

Consenting patients were registered and scheduled for an interview, usually on the day of their next appointment at the oncology clinic. Utilities were elicited directly from subjects about their current HRQL by one trained researcher using a standardised, face-to-face, time trade-off (TTO) interview with a hypothetical survival time of two years (*as described in section 3.2 of chapter 3*). The TTO was expressed on the standard continuum where 1 represents full health and 0 represents dead [16]. Patients were mailed the questionnaire and asked to complete it 3 to 7 days before the planned interview.

5.3.2 The Utility-Based Questionnaire-Cancer (UBQ-C)

The Utility-Based Questionnaire (UBQ) is a validated, disease-specific HRQL questionnaire that was designed to be an outcome measure for clinical trials in cancer and cardiovascular disease [12-13, 129]. The cancer version (the Utility-Based Questionnaire-Cancer, UBQ-C) includes 29 items about specific aspects of HRQL, and a global scale called the health status thermometer, which is a single item that asks respondents for a unified assessment of their health status. The 29 items about specific aspects of HRQL are grouped into subscales for physical function (3 items), social/usual activities (4 items), self-care (1 item), and distresses (21 items) due to physical and psychological symptoms associated with cancer and its treatment (*Reproduced in appendix 1*).

The UBQ-C was designed for use in clinical trials of cancer therapy, so it needed to be relevant to cancer patients, relatively brief and easy to self-complete. The form and content of the questionnaire builds on the conceptual framework for health status assessment developed by Gudex et al for an existing generic utility-based instrument called the Health Measurement Questionnaire (HMQ) [118]. The HMQ includes 36 items covering 5 key dimensions of HRQL (general mobility, usual activities, self-care activities, social and personal relationships, and psychological distresses). A generic core set of items from the HMQ was taken for the UBQ-C. Items less relevant for cancer patients (eg hearing, vision, writing, speaking and incontinence) were discarded, and the response formats of some items were modified. Cancer-specific items were selected for addition by a review of existing literature on HRQL instruments used in cancer patients [130-131]. Two measures of global health status were also added. The health status thermometer is similar to the graduated, vertical, visual analog scale that accompanies the EuroQol EQ-5D questionnaire [104-105], but with the anchors of 'best imaginable health state' and 'worst imaginable health state' replaced by 'full health' and 'death', to conform with the requirements of a utility scale. The UBQ-C also includes the general health item from the Short-Form-36 health survey (SF-36) [132], which is a widely used and extensively validated measure of generic health status [5]. This item asks respondents to 'describe their general health' as excellent, very good, good, fair or poor.

The psychometric properties of the UBQ-C in a cancer population have been reported previously [12-13]. These include good feasibility (high completion rate with little missing data), internal consistency of subscales (Cronbach's alphas > 0.75 and confirmatory factor analysis), test-retest reliability (intraclass correlation coefficients: median 0.85, lower quartile 0.81, upper quartile 0.90), convergent validity (substantial correlations with related instruments: GLQ-8, GLQ-Uniscale [130], Priestman and Baum LASA scales [131], and Life Satisfaction Index-A [133]), discriminative ability (between groups with different disease severity) and responsiveness to change within individuals.

5.3.3 Statistical methods

The scoring algorithm was produced by modelling the valuation survey data using the multi-step approach developed by Lumley et al [32]. A 'subset index' is calculated by combining the questionnaire subscales into a subset index according to weights based on their correlations with a global scale. Here we define a global scale as a single item asking respondents directly for a unified assessment of their HRQL [31]. An 'overall HRQL index' is then calculated by combining the subset index with this global scale using weights based on their statistical precision. Finally, a 'utility index' is calculated by transforming the overall HRQL index. A novel feature of Lumley's approach that is not incorporated in other approaches to deriving scoring algorithms [9, 39, 90] is that it combines a single-item global scale with multi-item subscales for specific aspects of HRQL. The purpose of including the global scale in the index is to incorporate information about any additional aspects of HRQL that are important but not captured by the subset index [32].

The scoring algorithm for the UBQ-C was derived in four steps (*figure 4.2 of chapter 4*). First, subscale scores for physical function, social/usual activities, self-care and distresses were calculated from the ratings on the relevant UBQ-C items. Second, a subset index was calculated by weighted combination of the subscale scores. Third, an overall HRQL index was calculated by weighted combination of the subset index and the health status thermometer. Fourth, the overall HRQL index was converted to a utility-based index with a suitable transformation. The following paragraphs describe each step in detail.

The subscale scores for physical function, social/usual activities, self-care and distresses are the simple averages of the relevant, non-missing items, linearly transformed to a scale from 0 (worst) to 1 (best). Responses to the items about 'Sex life' and 'Other problems' are not included when calculating the scores for the subscales because they are commonly omitted by respondents.

The subset index was calculated by weighted combination of the subscales for physical function, social/usual activities, self-care and distresses. Weights for the subscales (W1-4 in *figure 4.2 of chapter 4*) were derived from, and proportional to, the coefficients obtained from multivariable ordinary least squares linear regression

of the health status thermometer on the subscales. The weights are designed to reflect the relative contribution of each subscale to overall HRQL. The scores for the subset index for each subject were calculated by applying the weights to the subscale scores as follows:

[1] Subset index = [W1 * PF] + [W2 * SA] + [W3 * SC] + [W4 * DI]

W1-4 are the weights for the subscales: PF is physical function, SA is social/usual activities, SC is self-care, DI is distresses.

The score for the subset index was recorded as missing if any of its component scores were missing.

The overall HRQL index was calculated by weighted combination of the subset index with the health status thermometer. Greater weight was given to the component with least measurement error. The weights were calculated using Lumley's formulae, as follows:

[2] W = Var(T) * [1 – r(T)] / MSE(R)

W is the weight allocated to the subset index, so 1 – W is the weight allocated to the health status thermometer (*figure 4.2 of chapter 4*). Var(T) is the variance of the health status thermometer obtained from the dataset. r(T) is the intraclass correlation coefficient of the health status thermometer, and was calculated with test-retest data from a previous validation study [13]. MSE(R) is the mean square for error from the linear regression of the health status thermometer on the four subscales, and was obtained from the dataset.

The scores for the overall HRQL index for each subject were calculated as follows:

[3] Overall HRQL index = [W * Subset index] + [(1 – W) * HST]

HST is the health status thermometer. The scores for the overall HRQL index were recorded as missing if the score for the subset index or health status thermometer was missing.

A suitable transformation function was sought to convert the overall HRQL index to the utility index. We considered a range of functional forms used to transform measures of HRQL to measures of utility in previous studies [92]. We selected the

function that best mapped the relationship between the overall HRQL index and TTO utility in the development dataset. The details are described in the supplement to this chapter (*section 5.6*). The scores for the utility index for each subject were calculated by applying the chosen transformation function to the scores for the overall HRQL index.

Preliminary validation of the algorithm was done by comparing the scores on the utility index to those from other measures of HRQL, health status and utility. We assessed how closely the utility index was related to the TTO utility using Spearman's rank-order correlation (r_s) and paired t-tests. Associations between the utility index and two independent global measures of HRQL, the general health item from the SF-36 (referred to above) and the Spitzer-Uniscale of global life quality (*described in section 3.3 of chapter 3 and appendix 2*) [14-15], were also assessed with Spearman's rank-order correlation. We tested the hypothesis that compared with related global measures, the derived indices would give estimates of differences in mean scores between subjects that were grouped by their response to the general health item that were more precise (narrower confidence intervals) but unbiased (similar point estimate). The overall HRQL index was compared with the health status thermometer, and the utility index was compared with the TTO utility. Differences in mean scores between groups were calculated using unpaired t-tests. The relative precisions of the related measures were compared using the relative efficiency statistic [138, 144]. The reciprocal of the relative efficiency statistic is the factor by which the sample size can be reduced when a more precise and therefore more efficient scale is used. The relative efficiency statistic was calculated as the squared ratio of the t-score from the comparison of the groups using the derived index divided by the t-score from the comparison of the groups using the related global measure.

5.4 Results

The study profile describing the subjects and data used to generate the scoring algorithm is shown in *figure 3.2 of chapter 3*. Of the 323 patients that were approached to take part in the study, 204 were eligible. Compliance was excellent, with planned interviews and questionnaires completed by 98% of participants. All items on the UBQ-C except 'Sex life' and 'Other problems' were completed by over 90% of patients. Characteristics of eligible subjects are summarised in table *5.1*. All patients had advanced cancer, mostly arising from the breast or bowel. The mean age was 56 and most age groups were represented. The different levels of general health status were also well-represented. Most modes of treatment were represented: chemotherapy, supportive care and observation.

UBQ-C ratings on the health status thermometer and subscales are summarised in table *5.2*. Patients reported worst impairment for physical function and least impairment for self-care.

The derived weights for the subscales (W1-4) health status thermometer (1-W), and subset index (W) are shown in table *5.3*. The health status thermometer accounted for about two-thirds of the index for overall HRQL. Of the subscales, greatest weight was given to distresses and least to self-care.

Table 5.1 Valuation survey: patient characteristics

	(n=204)
	%
Cancer type	
Breast	**50**
Bowel	**29**
Other	**21**
Age (Years)	
< 40	**12**
40-49	**21**
50-59	**25**
60-69	**28**
≥ 70	**14**
Gender	
Male	**32**
Female	**68**
Marital status	
Partner	**65**
No partner	**35**
(single, divorced, widowed)	
Education	
Primary school	**4**
Some high school	**22**
Completed high school	**35**
Higher education	**39**
Country of origin	
Australia	**80**
Other	**20**
General health	
Excellent	**10**
Good	**48**
Fair	**35**
Poor	**7**

Table 5.2 Valuation survey: ratings on UBQ-C

	Mean	SD
Health status thermometer	**0.74**	**0.16**
Physical function	**0.65**	**0.23**
Social/usual activities	**0.77**	**0.22**
Self-care	**0.95**	**0.14**
Distresses	**0.80**	**0.15**

SD, standard deviation. UBQ-C, Utility-Based Questionnaire-Cancer. All ratings on scale from best (one) to worst (zero).

Table 5.3 Valuation survey: weights for scoring algorithm

		Weight
W	**Health status thermometer**	**0.65**
1-W	**Subset index**	**0.35**
W1	**Physical function**	**0.28**
W2	**Social/usual activities**	**0.06**
W3	**Self-care**	**0.01**
W4	**Distresses**	**0.64**

W is the weight allocated to the subset index. 1 – W is the weight allocated to the health status thermometer. W1-4 are the weights for the subscales.

The overall HRQL index was calculated by applying these weights to the subjects' ratings on the UBQ-C using formulae [1] and [3]. The transformation that best reflected the relationship between the overall HRQL index and TTO utility was a disutility power transformation (*supplementary section 5.6*), viz:

[4] Utility index = $1 - (1 - \text{overall HRQL index})^{2.03}$

This transformation was used to convert the overall HRQL index into the utility index.

Scores for the overall HRQL index, utility index and TTO utility are compared in table *5.4*. The TTO utility was 1.0 for about half the subjects, despite significant impairments in HRQL. Because of this skewed distribution, the mean value of the TTO utility was lower than its median value. There were no associations between the TTO utility and the patient characteristics listed in table *5.1* above (data not shown). The overall HRQL index gave substantially lower scores than the TTO utility (means 0.74 vs 0.92, difference 0.17, 95% CI 0.14 to 0.19). Scores were similar for the utility index and the TTO utility (means 0.92 vs 0.91, difference 0.01, 95% CI -0.02 to 0.03).

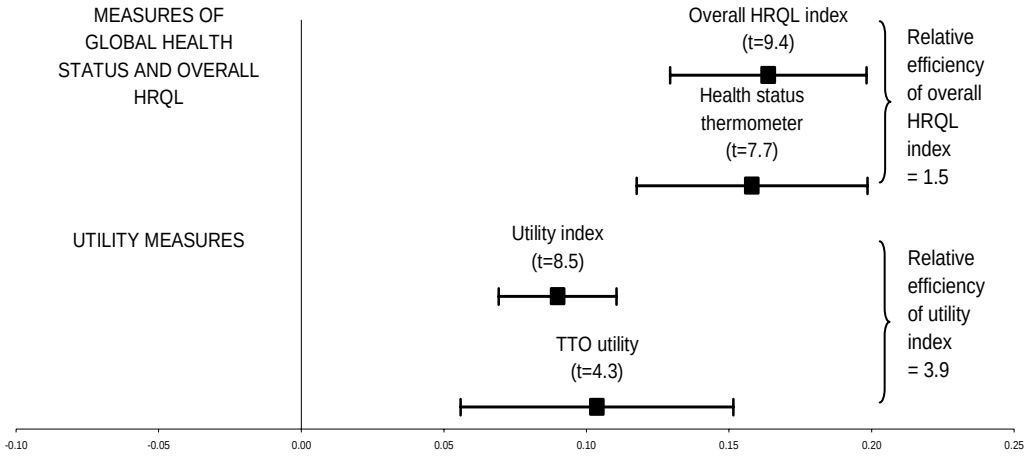
Table 5.4 Valuation survey: comparison of scores for overall HRQL index, utility index and time trade-off utility

Statistic	Time trade-off utility	Overall HRQL index	Utility index
Mean	**0.91**	**0.74**	**0.92**
Standard deviation	**0.17**	**0.14**	**0.08**
95% confidence intervals	**(0.89, 0.94)**	**(0.72, 0.76)**	**(0.90, 0.93)**
Median	**1.00**	**0.77**	**0.95**
Inter-quartile range	**(0.88, 1.00)**	**(0.63, 0.86)**	**(0.87, 0.98)**
% with score of 1.0	**50**	**0.5**	**0.5**
Mean difference compared to time trade-off utility	**0**	**0.17**	**0.01**
(p-value)	**N/A**	**< 0.0001**	**0.6**

HRQL, health-related quality of life. All ratings on scale from worst (0) to best (1).

Comparisons of the utility index with other measures of HRQL, utility and health status provide preliminary evidence of its validity. The utility index was moderately correlated with the TTO utility (r_s 0.38), the general health status item from the SF-36 (r_s 0.63), and the Spitzer-Uniscale of global life quality (r_s 0.68). The estimated differences in mean scores between subjects grouped by general health in the development dataset were more precisely estimated by the derived indices than by the health status thermometer or the TTO utility (figure *5.1*). The relative efficiency statistics in figure *5.1* correspond with reductions in sample size needed to detect a significant difference by using the indices of 33% for the overall HRQL index compared with the health status thermometer, and of 75% for the utility index compared with the TTO utility.

Figure 5.1 Valuation survey: comparison of precision of (i) overall HRQL index and health status thermometer, (ii) utility index and time trade-off utility, in distinguishing subjects grouped by their general health status (excellent or good versus fair or poor)



HRQL, Health-related quality of life. Relative efficiency, reciprocal of factor by which sample size can be reduced when more efficient index is used (see text). t, t-score for difference between groups. All ratings on scale from 0 to 1.

5.5 Discussion

We have derived a scoring algorithm for a disease-specific utility-based instrument that is designed to inform clinical decisions about cancer treatments. The algorithm converts ratings from a cancer-specific HRQL questionnaire called the UBQ-C into a utility-based index. Firstly, the algorithm calculates a subset index from a weighted combination of the UBQ-C subscales for physical function, social/usual activities, self-care and distresses. Secondly, an overall HRQL index is calculated from a weighted combination of the health status thermometer and the subset index. Thirdly, the algorithm calculates a utility index by applying a power transformation to the overall HRQL index. The scoring algorithm was developed using TTO utilities and UBQ-C ratings elicited from patients with advanced cancer who rated their current health status and HRQL. The utilities can be used to generate QALYs to compare cancer treatments.

Utilities and QALYs are a useful way to compare cancer treatments because they can be evaluated on a common scale that incorporates disparate treatment effects like gains in survival duration, improvements in HRQL due to relief of cancer symptoms, and deteriorations in HRQL due to treatment-related side effects [9, 75]. Analyses of cancer trials in terms of utilities and QALYs are increasingly used to inform economic decisions about cancer treatments [76-83], but can also be used to inform clinical decisions [60, 84-88]. Despite the advantages of utilities and QALYs, there is no standardised approach for eliciting utilities [9, 16, 147-148]. One way to obtain utilities is to elicit them directly from respondents using a standard gamble or time trade-off (TTO) interview, but this task is complex, resource intensive, and can be distressing if cancer patients are required to assign utilities for their own health state [16]. A more practical approach is to derive utility scores from a utility-based instrument. We posit that deriving utility scores from a utility-based instrument that is disease-specific and based on the perspective of patients is the best approach for informing clinical decisions.

The UBQ-C is a disease-specific instrument that is designed for the evaluation of cancer treatments. It asks about important consequences of cancer and its treatment not covered by generic instruments such as such as the EQ-5D [103], HUI3 [106] or

SF-6D [107] including fatigue, nausea, shortness of breath, and hair loss. The main advantage of a disease-specific instrument such as the UBQ-C over a generic instrument for generating utility scores is that it should be more sensitive to detect differences in health-related quality of life between individuals with cancer. This requires empirical testing, as has been done for other disease-specific instruments that generate utility scores for cancer [95, 98, 110] and a range of other diseases including bladder disorders [97, 113], hearing impairment [114] and asthma [111]. Another advantage of using a disease-specific utility-based instrument is that it provides data on specific aspects of HRQL, overall HRQL, and utility with a single questionnaire and increases the availability of utility data for comparisons of treatment from randomized clinical trials [98]. This approach enables utilities to be derived from previous studies where the UBQ-C was used, and reduces questionnaire burden for future trial participants by having a single questionnaire and approach that provides these 3 kinds of information.

The major limitation of using disease-specific, utility-based instruments is that the utility scores they provide may not be comparable to those derived from other instruments, particularly generic instruments, because the dimensions of health status and HRQL that they cover are different [68, 108-109]. Whether this is a problem depends on the decision for which the utilities are being applied. We argue that disease-specific instruments are best suited to treatment comparisons within a particular disease used to inform clinical decisions. In this context comparisons across other diseases and healthy populations are less important, but coverage of aspects relevant to the patients under study is crucial. Others have argued that disease-specific instruments may also be suitable for treatment comparisons across all diseases to inform health funding and policy decisions if the scoring algorithm is derived using a valuation technique and population sample that is similar to a generic instrument, and the utility scores are shown to be comparable [97].

The algorithm described in this study was based on the perspective of cancer patients who were currently experiencing those health states. The perspective differs in two important ways from scoring algorithms used for most of the generic and cancer-specific utility-based instruments reported previously. First, it is the perspective of patients rather than lay people. Second, it reflects views about a health state that is

real and current rather than hypothetical and in the future [149]. The perspective from which a utility is elicited may have significant implications for clinical and economic decisions that incorporate utilities and QALYs, because patients typically assign higher utilities to a given health state than lay people [56-58]. This may reflect partly the lay person's difficulty appreciating what a hypothetical health state is really like, and partly the patient's adaptation to their own health state [16, 58, 60-62]. The choice of perspective should reflect the viewpoint from which the results will be interpreted [16, 61]. As discussed in the introduction, it is generally agreed that the perspective of patients is more appropriate for informing clinical decisions about specific treatments, while the perspective of lay people is more appropriate for informing decisions about health policy and funding. Some also argue that the perspective of patients should be used to inform funding and policy decisions, because patients better understand what it is like to live with a particular disease [56, 59], but this argument is controversial because it runs counter to prevailing health economic guidelines [69-71].

This study also provides preliminary evidence supporting the validity of the utility index. It was substantially correlated with independent measures of general health, overall life quality, and TTO utilities. Mean scores for groups from the utility index and TTO utility were almost identical. This supports the validity of using the utility index to generate mean utilities for comparing patient groups. However, as expected we found that the utility index did not accurately predict utilities for individual patients. The mean absolute difference between the utility index and TTO utility for each subject was relatively large at 0.10. This finding argues against using the utility index to predict utilities for individuals. This is exactly as expected [98], because utilities are influenced by factors apart from HRQL such as individuals' attitudes to risk and uncertainty (for the standard gamble), discount rate (for the time trade-off), and idiosyncratic preferences) [16, 148].

The derived indices for HRQL and utility gave more precise estimates of differences between groups than the health status thermometer or TTO utility. We expected more precise estimates because any score aggregated from multiple items will produce a more precise estimate of differences between groups than a single-item scale [22, 32]. This finding does not strengthen or weaken the validity of the indices but is an

expected measurement property which enhances the sensitivity and responsiveness of the indices.

Ongoing work is needed to support the validity of the utility index. A companion paper describes the application of the scoring algorithm to independent trial datasets in breast cancer, and provides further evidence to support its validity by comparison with clinical data [146] (*Chapter 6*). We have also reported on a comparison of the value and sensitivity of utility scores generated by the index to those generated by the EQ-5D in colorectal cancer [150]. Independent testing in other datasets will further establish validity.

The study population and valuation survey used to develop the scoring algorithm has several strengths. The patient characteristics were diverse including men and women with a broad range of ages, levels of performance status, and levels of health status. Compliance was excellent with both UBQ-C completion and utility interviews. We used the TTO method to elicit utilities for health states. The TTO is practical, reliable and has empirical validity [16, 49, 151]. A limitation of our valuation survey was that its sample size was too small to allow division of the group into a 'training' set, where the algorithm was developed, and a 'validation' dataset where its validity and accuracy was independently tested. The dataset was confined to patients with advanced cancer, mostly with breast or colorectal primaries, and attending ambulatory clinics. This may raise questions about the generalisability of the algorithm and approach to patients with cancers that are of earlier-stage, in remission, or from other primary sites.

The novelty of our statistical approach is in its combination of a single-item global scale with multi-item subscales for specific aspects of HRQL and its methods for deriving optimal weights. Most other utility-based indices do not incorporate a single-item global scale [9, 39, 90]. Incorporation of the single-item global scale has two potential advantages. First, it provides a unified reflection of how the patient rates their health status that enables incorporation of aspects of HRQL that are important but are not directly captured by multi-item subscales [32]. Second, it allows the scoring algorithm to be optimised in different treatment contexts by adjusting the weights assigned to the multi-item subscales [32]. The purpose of

optimising the algorithm is to reflect the differences in importance that patients with different types and stages of cancer, and treatments assign to various dimensions of HRQL [32, 63]. The implications of optimising the algorithm for different treatment contexts are addressed by application and discussion in a companion paper [146] (*Chapter 6*).

This work enables HRQL data obtained with a simple cancer-specific questionnaire to be converted into a utility index that reflects the perspective of cancer patients. The approach is best-suited to generating estimates of mean utilities for groups, and our work so far supports this application. It can be applied in clinical trials to compare the effect of cancer treatments on HRQL using utility measures, and to generate QALYs for informing clinical decisions and as an alternate viewpoint for economic analyses. The approach provides a general method for converting HRQL ratings to valid utility-based measures that could be applied in other trial settings for analysis of HRQL data collected with different questionnaires.

5.6 Supplementary section

Here we describe how we derived an equation to convert the overall HRQL index to the utility index. The scores on the overall HRQL index are expressed with a 'value-based' scaling method, and the scores on the utility index are expressed with a 'utility-based' scaling method. We define these terms in the following paragraphs based on descriptions by McDowell et al [5].

We refer to a measure expressed with a value-based scaling method as one that measures a respondent's perceptions about the presence and severity of symptoms or disabilities [5]. The value-based scaling method is sometimes referred to as the psychometric scaling method because the numerical scores are assigned to the responses in a way that is derived from the psychometric tradition. The psychometric tradition focuses on perceptions or feelings and usually refers to current health status rather than future outcomes [5]. The responses to the individual items on the Utility-Based Questionnaire-Cancer (UBQ-C) and the scores on the overall HRQL index that are derived from them are expressed with a value-based scaling method, as is the 'health status thermometer' that assigns a perception of full health a score of 100 and death 0.

In contrast, we refer to a measure expressed with a utility-based scaling method as one that measures a respondent's strength of preference for particular outcomes when faced with uncertainty such as the possibility of a future gain [5, 9, 16]. The utility-based scaling method is sometimes referred to as the econometric scaling method because the numerical scores are assigned to the responses in a way that is derived from the econometric tradition. The econometric tradition focuses on strength of preference for alternate outcomes that will occur in the future. The responses to a time trade-off interview are expressed with a utility-based scaling method, as is a utility index that is derived from a utility-based instrument.

The value-based scaling method records values and the utility-based scaling method records utilities. The distinction between utilities and values is important because utilities have scaling properties that are designed for generating quality-adjusted life-years (QALYs). This is because utilities capture both the person's preference and

their attitude towards risk, which is relevant for studies of choices between alternative therapies for which the outcome lies in the future and remains uncertain [5]. The utility-based scaling method was chosen as the basis for expressing quality of life because it offers a way to integrate morbidity and mortality into a single scale called quality-adjusted life-years (QALYs) [5].

Because of the complexity of the utility-based scaling method, methods have been developed to convert measures expressed with a value-based scaling method (referred to in this section as 'Q') to measures expressed with a utility-based scaling method (referred to in this section as 'U'). Several functions have been used in other studies to transform Q into U (see [37] for an overview). The functions vary in their form, complexity, and rationale; and include simple linear [152-155], quadratic [156-157], cubic [156-157], plateau models [154], and power transformations [51, 93, 106, 158-162]. Some directly transform Q to U, others transform impairment in quality (1-Q) to disutility (1-U). There is no consensus on the best functional form [37, 157], but there are two main criteria. One is that the functional form has an underlying conceptual rationale. The other is that it has predictive precision, which tests if the function predicts a value of U from Q that is similar to the true value of U.

The power transformation is the best studied function. The utility power transformation is of the form $U = Q^k$, and the disutility power transformation is of the form $(1-U)=(1-Q)^k$ [92]. An advantage of the power transformation is that the value of U will be between zero and one for any value of Q that is between zero and one, which satisfies one requirement of the utility-based scaling method [16]. The rationale for a power transformation is that Q is expressed with a value-based scaling method under conditions of certainty and lack of choice, and U is expressed with a utility-based scaling method under conditions of uncertainty and choice, so the difference between Q and U is due to the degree of risk aversion that subjects have about making choices with uncertain outcomes [163]. The exponent *k* adjusts Q for the degree of risk aversion, and its value varies between studies [92]. The predictive precision of the power transformation varies in different contexts. It appears better when three criteria are satisfied: U and Q are elicited from lay people rather than patients; the aggregate (mean or median) values for U and Q are calculated for each health state; and the test of predictive precision is based on the relationship between

values of U and Q at the aggregate level rather than at the individual respondent level. The predictive precision of the power transformation appears lower when these three criteria are not satisfied [37].

Four other functions have been used to convert Q into U. The most parsimonious is a simple linear model: $U = b_0 + b_1 Q$. A limitation of the simple linear model is that values of U are not bound between zero and one. More complex functions are disutility functions with no intercept: $1 – U = b_1 (1 - Q)$; quadratic functions with no intercept: $U = Q^2 + b_1 (Q – Q^2)$; and cubic functions with no intercept: $U = Q^3 + b_1 (Q – Q^3) + b_2 (Q^2 – Q^3)$ [157, 164]. A problem with these functions is the lack of an underlying conceptual rationale [37]. The more complex cubic and quadratic models have better predictive precision than the linear models. The cubic and quadratic models have better predictive precision than the power transformation in some studies [157] but not in others [156].

We selected the function that best described the relationship between the overall HRQL index and the TTO utility in the valuation survey. As described previously, the measures were moderately correlated (Spearman correlation 0.38), however the overall HRQL index was typically substantially lower than the TTO utility (mean difference 0.17, 95% CI 0.14 to 0.19, $p < 0.0001$) (*table 5.5, figure 5.2*).

A limitation of each of the six tested functional forms is that each ignored the half of respondents who assigned a utility of 1. The power transformation assigned a utility of less than 1 to any respondent with any impairment in HRQL at all. Future work could compare the power transformation to a plateau model. This would model a linear relationship where utilities greater than 1 are assigned a utility of 1 [154].

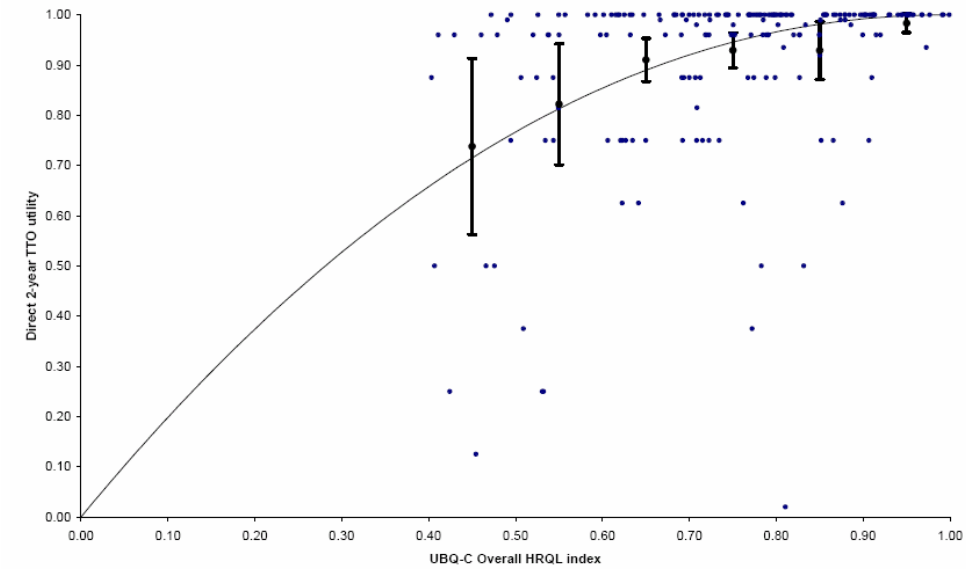
We considered six potential functional forms which are outlined in table *5.5*. Optimal parameters for each of the functional forms were estimated by least squares regression of the TTO utilities on the overall HRQL index. Our primary criterion for selecting a function was predictive precision, which was compared in terms of mean absolute error, root mean square error, and proportion of predicted values within 0.05 of actual values.

Table 5.5 Functions to convert overall HRQL index to utility index

N°	Function	Parameters		Predictive precision			
		Parameter estimates	95% CI	Root mean square error	Mean absolute error	Errors < 0.05 (%)	Errors < 0.1 (%)
1	Utility power transformation: $U = Q^k$	k=0.33	0.25, 0.41	0.16	0.11	36	65
2	Disutility power transformation: $(1-U) = (1-Q)^k$	k=2.03	1.78, 2.28	0.16	0.10	52	68
3	Utility linear function: $U = b_0 + b_1 Q$	b_0=0.59 b_1=0.43	0.46, 0.71 0.26, 0.59	0.16	0.11	36	64
4	Disutility linear function with no intercept: $1 – U = b_1 (1, Q)$	b_1=0.38	0.30, 0.47	0.16	0.11	33	65
5	Quadratic function with no intercept: $U = Q^2 + b_1 (Q – Q^2)$	b_1=1.98	1.85, 2.12	0.16	0.10	50	67
6	Cubic function with no intercept: $U = Q^3 + b_1 (Q – Q^3) + b_2 (Q^2 – Q^3)$	b_1= 2.59 b_2=-2.49	1.88, 3.29 -4.22, 0.76	0.16	0.10	38	68

Abbreviations, CI, confidence interval. ‘Errors < 0.05’ and ‘Errors < 0.1’, proportion of subjects where difference in value between utility predicted by function and directly elicited utility is < 0.05 or 0.1. Q, overall HRQL index. U, utility index.

Figure 5.2 Valuation survey: relationship of overall HRQL index and time trade-off utility



Scatter plot of time trade-off utility versus overall HRQL index.

Superimposed is box and whiskers of time trade-off utility (Mean, 95% CI) grouped by overall HRQL index categories (cutpoints: 0.5, 0.6, 0.7, 0.8, 0.9).

Superimposed is regression line of power transformation ie. utility index = 1 - (1 – overall HRQL index)$^{2.03}$.

We also checked for violations of the statistical assumptions underlying these functions by testing for bias of the mean error. Analyses were performed using the PROC REG and PROC NLIN functions of SAS System for Windows Release 8.02 [145].

The predictive precision of the six tested functional forms is compared in table *5.5* above. The disutility power transformation was selected because it had the best predictive precision on all criteria. The exponent for the power transformation was similar to that found in other studies [92]. The power transformation performed better in our dataset than more complex quadratic and cubic functions that others have advocated [157, 164] when we used similar selection criteria [93, 164]. We converted the overall HRQL index to the utility index with this disutility power transformation. There was no systematic bias in errors within its regression model (mean -0.01, 95% CI -0.03 to 0.01, p=0.3). When patients were grouped by their general health, mean scores on the utility index and TTO utilities were similar (table *5.6*). The proportion of scores on the utility index within 0.1 and 0.05 of TTO utilities was better for subjects with general health that was excellent or good compared to fair or poor (table *5.6*).

In conclusion, a disutility power transformation was used to convert the overall HRQL index to the utility index. This is a standard method to convert a measure from a value scale to a utility scale. The equation has adequate predictive precision to compare utility scores between groups in clinical trials.

Table 5.6 Valuation survey: comparison of utility index and TTO by general health status

Group	%	Utility index		Time trade-off utility		Errors < 0.05
		Mean	95% CI	Mean	95% CI	%
All subjects	100	0.92	0.90, 0.93	0.91	0.89, 0.94	52
SF-36 general health						
Excellent	10	0.99	0.98, 1.0	0.99	0.98, 1.0	100
Good	48	0.95	0.94, 0.96	0.95	0.92, 0.97	65
Fair	35	0.89	0.87, 0.91	0.88	0.83, 0.92	31
Poor	7	0.81	0.74, 0.87	0.73	0.57, 0.89	8

SF-36 General health, response to Short-Form-36 general health item; Errors < 0.05 (%), proportion of patients where absolute difference between utility index and directly elicited utility is < 0.05.

6. Preliminary validation of an optimally-weighted patient-based utility index by application to randomised trials in breast cancer

6.1 Overview

This chapter is a published work. The entire manuscript is quoted verbatim, and amendments are presented in italics.

Publication details

Grimison PS, Simes RJ, Hudson HM, Stockler MR.
Preliminary validation of an optimally-weighted patient-based utility index by application to randomised trials in breast cancer. Value in Health 2009; 12(6): 967-976 (PMID 10490566).

Contribution of authors

PSG developed the research proposal, selected the research methods, did data analysis, interpreted the findings, and drafted the manuscript.

RJS conceived the research proposal, participated in selection of research methods and interpretation of findings, and contributed to the drafting and revision of the manuscript.

HMH provided guidance for data analysis and interpretation of the findings, and contributed to the drafting and revision of the manuscript.

MRS contributed to the conception and development of the research proposal, selection of research methods, interpretation of findings, and drafting and revision of the manuscript.

Abstract

Objectives: To optimise, apply and validate a scoring algorithm that provides a utility index from a cancer-specific quality of life questionnaire called the Utility-Based-Questionnaire-Cancer (UBQ-C) using datasets from randomised trials in breast cancer. The index is designed to reflect the perspective of cancer patients in a specific clinical context so as to best inform clinical decisions.

Methods: We applied the UBQ-C scoring algorithm to trials of chemotherapy for advanced (n=325) and early (n=126) breast cancer. The algorithm converts UBQ-C subscales into a subset index, and combines it with a global health status item into an overall HRQL index, which is then converted to a utility index using a power transformation. The optimal subscale weights were determined by their correlations with the global scale in the relevant dataset. The validity of the utility index was tested against other patient characteristics.

Results: Optimal weights (range 0-1) for the subset index in advanced (early) breast cancer were: physical function 0.20 (0.09), social/usual activities 0.23 (0.25), self-care 0.04 (0.01), distresses 0.53 (0.64). Weights for the overall HRQL index were health status 0.66 (0.63) and subset index 0.34 (0.37). The utility index discriminated between breast cancer that was advanced rather than early (means 0.88 vs 0.94, $p<0.0001$) and was responsive to toxic effects of chemotherapy in early breast cancer (mean change 0.07, $p<0.0001$).

Conclusions: The scoring algorithm for the UBQ-C utility index can be optimised in different clinical contexts to reflect the relative importance of different aspects of quality of life to the patients in a trial. It can be used to generate sensitive and responsive utility scores, and quality-adjusted life-years, that can be used within a trial to compare the net benefit of treatments and inform clinical decision-making.

6.2 Introduction

The quality-adjusted life-year (QALY) approach is a useful way to compare cancer treatments, because it integrates the beneficial and harmful effects of treatment on health-related quality of life (HRQL), expressed as a utility, with the effects of treatment on survival [9, 61, 75]. Analyses of cancer trials in terms of utilities and QALYs are increasingly used to inform economic decisions about cancer treatments [76-81, 83], but can also be used to inform clinical decisions [60, 84-88].

A practical and feasible approach to obtain utility scores for generating QALYs in cancer trials is to use a utility-based instrument. A utility-based instrument uses a scoring algorithm to convert the responses from a questionnaire that elicits ratings about various dimensions of HRQL to a utility index [9, 16, 90]. The scoring algorithm is valued in a valuation survey, where a sample of people directly assign a utility score to the health states described by the questionnaire using a time trade-off interview or related technique [39]. A utility-based instrument may include generic or disease-specific questions, and the scoring algorithm may generate utilities that are based on the perspective of lay people or patients. Three of the most commonly used instruments are the EuroQol EQ-5D [103], Health Utilities Index (HUI3) [106] and SF-6D [107]. These instruments include generic questions applicable to any disease or population, and their scoring algorithms are based on the perspective of lay people. Utility-based instruments reported more recently have included disease-specific questions and use scoring algorithms that are based on the perspective of patients rather than lay people [97-98, 165].

Ideally, the perspective from which a utility instrument is valued should reflect the views of the population that the researcher is trying to reflect in the decision-making [9, 16, 36, 99]. In a companion paper we emphasised that patients typically assign a higher utility to a health state than a lay person, which can have significant implications for health funding, policy and clinical decisions that incorporate utilities and QALYs [165]. Researchers using utilities to inform health funding and policy decisions will generally prefer the perspective of lay people [69-71], while researchers using utilities to inform clinical decisions will generally prefer the perspective of patients [9, 16, 36, 62, 73]. This is because the objective of clinical

decisions is to maximise health for an individual patient with that disease [165]. Recently, it has been recognised that the preferences and attitudes of lay people in different countries may differ, because of differences in demographic background, social and cultural values, and political and economic systems [66-67]. As a result, some scoring algorithms for utility-based instruments based on the perspective of lay people have been optimised for use in different countries to reflect these differences [66, 100-102]. It has also been recognised that the preferences and attitudes of cancer patients in different clinical contexts may differ, because patients with different cancer diagnoses, stages of disease and treatment may assign different importance to different aspects of HRQL [32, 63-65]. We posit that scoring algorithms for utility-based instruments based on the perspective of patients should be optimised for different treatment contexts to reflect these differences.

Lumley et al have developed a novel approach to optimising scoring algorithms for different clinical contexts using the HRQL data collected in that context [32]. Lumley's approach requires a questionnaire including items about specific aspects of HRQL and a single-item global scale. We define a single-item global scale as one asking respondents directly for a unified assessment of their HRQL. Lumley's approach gives extra weight to the responses about specific aspects of HRQL that are more highly correlated with the responses on the global scale. These weights are intended to reflect the relative importance that the subjects assign to different aspects of HRQL. The optimisation of the scoring algorithm requires weighting to be determined for each clinical context but does not require the valuation survey to be repeated.

The aim of this work was to use Lumley's approach to derive an optimised scoring algorithm for a cancer-specific HRQL instrument that is based on the perspective of cancer patients. In a companion paper we described the development and preliminary validation of the algorithm [165] (*Chapter 5*). This paper describes the application of the algorithm to trial datasets, and illustrates how it can be optimised in different treatment contexts.

6.3 Methods

6.3.1 Sources of data

The data used to optimise, apply and validate the scoring algorithm were collected in two randomised clinical trials of chemotherapy for breast cancer. Both studies were approved by the human research ethics committees at all participating institutions. All patients provided written informed consent.

The first trial, referred to as the "advanced cancer trial", was conducted by the Australian New Zealand Breast Cancer Trials Group. It included patients with advanced breast cancer who were randomly allocated to receive either daily oral capecitabine or standard CMF as first-line chemotherapy until disease progression [124]. The primary outcome measure of the trial was quality-adjusted time to progression. Secondary outcome measures were time to progression, response rates, HRQL, overall survival, safety and cost-effectiveness. Eligible subjects were 18 years or older, and were about to start first-line chemotherapy for histologically confirmed advanced breast cancer. Subjects were excluded if they were totally confined to bed and completely disabled (ECOG performance status 4, as described in the next section). Enrolment was from June 2001 to July 2005 at 34 centres in Australia and New Zealand. Subjects completed the Utility-Based Questionnaire-Cancer (UBQ-C) and other questionnaires about HRQL that are described below (unless they could not read English). The data described in this paper come from baseline questionnaires competed prior to randomisation.

The second trial, referred to as the "early cancer trial", was conducted by the Australian New Zealand Breast Cancer Trials Group in collaboration with the International Breast Cancer Study Group. It included patients with high-risk early stage breast cancer who were randomly allocated to receive either high-dose chemotherapy with stem cell support over 12 weeks or standard-dose chemotherapy over 24 weeks [136]. The primary outcome measure of the trial was overall survival. Secondary outcome measures were quality-adjusted survival, disease-free survival, toxicity, HRQL and cost-effectiveness. Eligible subjects were aged 16 to 65 years, and were about to start adjuvant chemotherapy for histologically confirmed early-stage primary breast cancer with 5 or more involved axillary nodes. Subjects were

excluded if they were capable of only limited self-care and/or were confined to a bed or chair for more than 50% of waking hours (ECOG performance status 3 or 4). Enrolment was from March 1997 until March 2000 at multiple centres in Australia, New Zealand, Europe and Asia. Subjects living in Australia and New Zealand were eligible to participate in a substudy. Substudy participants were required to provide detailed information about HRQL and resource usage by completing the UBQ-C and other questionnaires described below. Questionnaires were completed prior to starting chemotherapy (baseline), 12 weeks after randomisation (during chemotherapy), and a few months after completing chemotherapy.

6.3.2 Questionnaires and other characteristics of subjects

The UBQ-C is a validated cancer-specific questionnaire that was designed as an outcome measure for clinical trials in the field of cancer. It includes 29 items about specific aspects of HRQL and a single-item global scale that asks respondents to rate their global health status (health status thermometer) [12-13, 165]. The 29 items about specific aspects of HRQL are grouped into subscales for physical function (3 items), social/usual activities (4 items), self-care (1 item) and distresses (21 items) due to physical and psychological symptoms relevant to cancer and its treatment. The UBQ-C also includes the general health item from the Short-Form-36 health survey (SF-36) [132]. More details about the conceptual framework, development, composition and psychometric properties of the UBQ-C are given in a companion paper [165] (*Chapter 5*).

Two additional questionnaires were completed. The Spitzer-Uniscale of global life quality was completed by all subjects as an additional global scale, but with the anchors of 'highest quality' and 'lowest quality' replaced by 'best possible' and 'worst possible' [14-15]. The Priestman and Baum Linear Analog Self Assessment Scales (LASAS) were completed by subjects in the advanced trial as validated measures of cancer-specific HRQL that include five scales about physical well-being, mood, pain, nausea and vomiting, and appetite [131, 134]. Clinicians completed the Eastern Cooperative Oncology Group (ECOG) performance status scale in the advanced trial. This rates patients' physical functional status as '0', fully active; '1', restricted in physical activity but able to do light work; '2', confined to a bed or chair for less than 50% of waking hours and capable of all self-care but unable

to do any work; '3', confined to a bed or chair for more than 50% of waking hours but capable of limited self-care; '4', totally confined to bed or chair, completely disabled, incapable of any self-care [135]. These questionnaires were described in *chapter 3* (*section 3.3*) and are presented in *appendix 2*.

6.3.3 Statistical methods

We optimised the scoring algorithm described in detail in a companion paper [165] and applied it to the clinical trial datasets. The scoring algorithm was outlined in figure *4.1 of chapter 4*.

Firstly, we calculated subscale scores for physical function, social/usual activities, self-care and distresses as the simple average of the non-missing items, linearly transformed to a scale from 0 (worst) to 1 (best).

Indices were then calculated by applying the following formulae:

[1] Subset index = [W1 * PF] + [W2 * SA] + [W3 * SC] + [W4 * DI]

[2] W = Var(T) * [1 – r(T)] / MSE(R)

[3] Overall HRQL index = [W * Subset index] + [(1 – W) * HST]

[4] Utility index = $1 - (1 - \text{overall HRQL index})^{2.03}$

W1-4 are the weights for the subscales, PF is physical function, SA is social/usual activities, SC is self-care, DI is distresses, HST is the health status thermometer. W is the weight allocated to the subset index, so 1 – W is the weight allocated to the health status thermometer. Var(T) is the variance of the health status thermometer obtained from the dataset. r(T) is the intraclass correlation coefficient of the health status thermometer, and was calculated with test-retest data from a previous validation study [13]. MSE(R) is the mean square for error from the linear regression of the health status thermometer on the four subscales, and was obtained from the dataset.

Optimal weights for the subscales (W1-4), subset index (W) and health status thermometer (1-W) were derived for each trial using the ratings on the UBQ-C in the relevant dataset. Weights W1-4 were derived from and proportional to the coefficients obtained from multivariable, ordinary least squares regression of the

health status thermometer on the subscales. Weights W and (1-W) were derived using formula [2] above.

The weights were then applied using formulae [1], [3] and [4] above to calculate scores for the subset index, overall HRQL index and utility index for each subject in each trial.

We examined the validity of the utility index against other characteristics of subjects. We tested its convergent validity, discriminative validity, responsiveness and predictive validity by comparing it with other self-rated measures of HRQL and with measures of physical function, cancer stage, treatment phase and subsequent survival.

Convergent validity tests how closely a measure is associated with related measures [2, 166]. The convergent validity of the utility index was tested by Spearman rank correlation (r_S) with the Spitzer-Uniscale, the SF-36 general health item, and scales from the Priestman and Baum LASAS questionnaire referred to above. We expected substantial correlations with the Spitzer-Uniscale and the SF-36 general health item. Three clinical experts made a priori hypothesis about the expected values of r_S with the LASAS scales as: insignificant (< 0.3), moderate (0.3-0.44), substantial (0.45-0.59), or high (> 0.6). Hypotheses were considered supported by the data if the observed r_S were at least as high as the median of the experts' expected r_S.

Discriminative validity tests how well a measure can distinguish between groups defined by an alternate criterion [28, 166]. The discriminative validity of the utility index was tested by its ability to detect cross-sectional differences between subjects with differing physical function as rated by their clinicians on the ECOG performance status scale referred to above. We also compared the discriminative ability of the UBQ-C overall HRQL index with that of the health status thermometer and the Spitzer-Uniscale. Differences between groups were evaluated with students t-test.

Responsiveness tests the ability of a measure to detect clinically important change over time [28, 144]. The responsiveness of the utility index was tested by comparing

scores in the early cancer trial before, during and after chemotherapy using paired t-tests.

Predictive validity tests how closely a measure is associated with a subsequent outcome [2, 166]. The predictive validity of the utility index was tested by its ability to predict survival duration in the advanced cancer trial, based on the hypothesis that overall survival in advanced cancer should be associated with baseline HRQL [167-171]. The strength of association between the utility index and survival duration was tested with the logrank test, by dichotomising subjects into a 'poor HRQL' group (utility index less than or equal to the median) and a 'good HRQL' group (utility index greater than the median).

The optimised scoring algorithm was applied to inform a specific treatment comparison of high-dose versus standard-dose chemotherapy for high-risk early-stage breast cancer using the data collected during chemotherapy from the early cancer trial. First, we compared scores on the utility index for participants allocated to each treatment arm using unpaired t-tests. Second, we used the index to reflect the relative importance of the effects of chemotherapy on different aspects of HRQL by comparing the weights allocated to each subscale. Third, we tested the hypothesis that the overall HRQL index compared with the health status thermometer would give an estimate of the difference in mean scores between treatment groups that was more precise but unbiased. The relative precisions of the related measures were compared using a measure called the relative efficiency statistic [138, 144]. The reciprocal of the relative efficiency statistic is the factor by which the sample size can be reduced when a more precise and therefore more efficient scale is used. The relative efficiency statistic was calculated as the squared ratio of the t-score for the index when comparing groups divided by the t-score for the related global measure when comparing groups.

6.4 Results

6.4.1 Study profiles and patient characteristics

The study profiles describing the subjects in each trial are shown in *figures 3.3 and 3.4 of chapter 3*. For the advanced cancer trial, compliance was excellent with questionnaires completed by 95% of subjects who were expected to complete them. For the early cancer trial, compliance was not as good with questionnaires completed by 72% prior to chemotherapy, 40% during chemotherapy, and 88% after completing it. All items on each UBQ-C questionnaire except for 'Sex life' and 'Other problems' were completed by over 90% of subjects in both trials. Characteristics of the 421 patients are shown in table *6.1*. Data was obtained from patients with breast cancer of both early and advanced stages. All subjects were female and most age groups were represented. For the advanced cancer trial, most had good performance status (ECOG 0 in 34% and ECOG 1 in 54%), and fewer had poor performance status (ECOG 2 in 11% and ECOG 3 in 2%). Ratings of general health ranged from 'Excellent' to 'Poor'.

Subjects' ratings on the UBQ-C are summarised in table *6.2*. At baseline, patients with advanced cancer reported worse health status than patients with early cancer as expected (means of 0.69 vs 0.81, difference 0.13 [with rounding], 95% CI 0.08 to 0.17, $p<0.0001$). Patients with early cancer reported worse health status during chemotherapy than before starting it (means 0.68 vs 0.81, mean deterioration 0.13, 95% CI 0.08 to 0.19, $p<0.0001$); or after finishing it (means 0.68 vs 0.84, mean improvement 0.15 [with rounding], 95% CI 0.10 to 0.21, $p<0.0001$). Similar differences were reported for ratings on UBQ-C subscales (table *6.2*).

Table 6.1 Breast cancer trials: patient characteristics

Dataset	Advanced cancer trial (n=325)	Early cancer trial (n=126)
Cancer stage	Advanced	High-risk early-stage
Cancer type (%)		
Breast	100	100
Gender (%)		
Female	100	100
Age (Years) (%)		
< 40	2	14
40-49	12	47
50-59	29	35
60-69	36	3
≥ 70	21	-

Dataset	Advanced cancer trial	Early cancer trial		
Treatment phase	Before treatment (n=295)	Before treatment (n=91)	During treatment (n=51)	After treatment (n=111)
General health (%)				
Excellent	6	22	6	22
Very good*	18	-	-	-
Good	30	54	40	66
Fair	32	19	42	9
Poor	13	4	12	3

* Response category 'Very good' not included in some versions of 'General health' item of UBQ-C

Table 6.2 Breast cancer trials: ratings on UBQ-C, overall HRQL index, and utility index

Dataset	Advanced cancer trial		Early cancer trial					
Treatment phase	Before treatment		Before treatment		During treatment		After treatment	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Health status thermometer	0.69	0.20	0.81	0.15	0.68	0.21	0.84	0.13
UBQ-C subscales								
Physical function	0.53	0.32	0.77	0.21	0.63	0.24	0.80	0.20
Social/usual activities	0.66	0.29	0.74	0.23	0.69	0.22	0.88	0.17
Self-care	0.89	0.20	0.89	0.15	0.97	0.10	0.99	0.06
Distresses	0.78	0.15	0.77	0.15	0.69	0.18	0.83	0.13
Overall HRQL index	0.69	0.18	0.80	0.13	0.68	0.18	0.84	0.12
Utility index	0.88	0.13	0.94	0.07	0.87	0.15	0.96	0.06

SD, standard deviation. UBQ-C, Utility-Based Questionnaire-Cancer. HRQL, health-related quality of life. All ratings on scale from best (one) to worst (zero).

6.4.2 Optimised scoring algorithms

The optimised index weights for the subset index (W), health status thermometer (1-W), and subscales (W1-4) for each trial are shown in table *6.3*. The weight assigned to the health status thermometer was similar for each trial and accounts for about two-thirds of the overall HRQL index. Of the subscales, greatest weight was given to distresses and least to self-care. The ordering of the weights assigned to the advanced cancer trial and early cancer trial were similar. Distresses were assigned the greatest weight, followed by social/usual activities, physical function, and self-care. However greater weight was assigned to distresses, and less weight to physical function and self-care, in women with early breast cancer than in women with advanced cancer.

Table 6.3 Breast cancer trials: weights for scoring algorithm

		Weights	
		Advanced cancer trial	Early cancer trial
W1	Physical function	0.20	0.09
W2	Social/usual activities	0.23	0.25
W3	Self-care	0.04	0.01
W4	Distresses	0.53	0.64
1-W	Health status thermometer	0.66	0.63
W	Subset index	0.34	0.37

1-W, W, W1-4 refer to the weights assigned to the health status thermometer, subset index, and subscales in formulae [1] and [2] (see text, and *figure 4.2 of chapter 4*)

6.4.3 Validation

Comparisons of the utility index with other characteristics of subjects supported its validity.

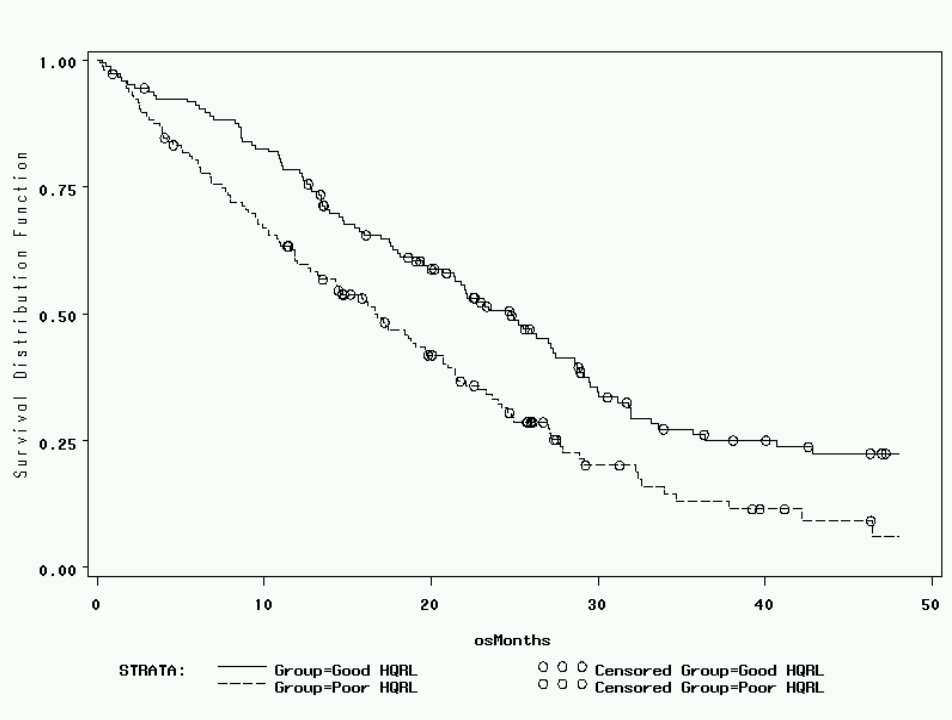
The convergent validity of the utility index was supported by its substantial correlation with the SF-36 general health item in both trials (r_S 0.74 in advanced and 0.64 in early) and the Spitzer-Uniscale in advanced cancer (r_S 0.71). There was also complete concordance of all expected and observed correlations of the utility index with the Priestman and Baum LASAS in the advanced cancer trial (data not shown).

The discriminative validity of the utility index was supported by strong evidence that subjects with early breast cancer prior to starting chemotherapy had higher utilities than those with advanced breast cancer (mean difference 0.07 [with rounding], 95% CI 0.04 to 0.10, p<0.0001) (table *6.2*). The discriminative validity of the utility index was also supported by its ability to distinguish subjects with differing performance status (PS) as rated by their clinicians in the advanced cancer trial (good performance status: mean 0.90, 95% CI 0.88 to 0.91; poor performance status: mean 0.73, 95% CI 0.67 to 0.79; mean difference 0.17, 95% CI 0.12 to 0.21; p < 0.0001).

The responsiveness of the utility index was supported by strong evidence that subjects with early breast cancer had higher utilities before starting chemotherapy than during it (mean difference 0.07, 95% CI 0.04 to 0.10; p<0.0001) (table *6.2*).

The predictive validity of the utility index was supported by its ability to predict survival duration in the advanced cancer trial when patients were divided into roughly equal-sized groups above and below the median score on the utility index (figure *6.1*). There was strong evidence that subjects with worse scores on the utility index at baseline (< 0.92) had shorter survival than those with higher scores (median 17 versus 23 months, log-rank p=0.005).

Figure 6.1 Advanced cancer trial: Kaplan-Meier plots for survival duration of subjects grouped by utility index

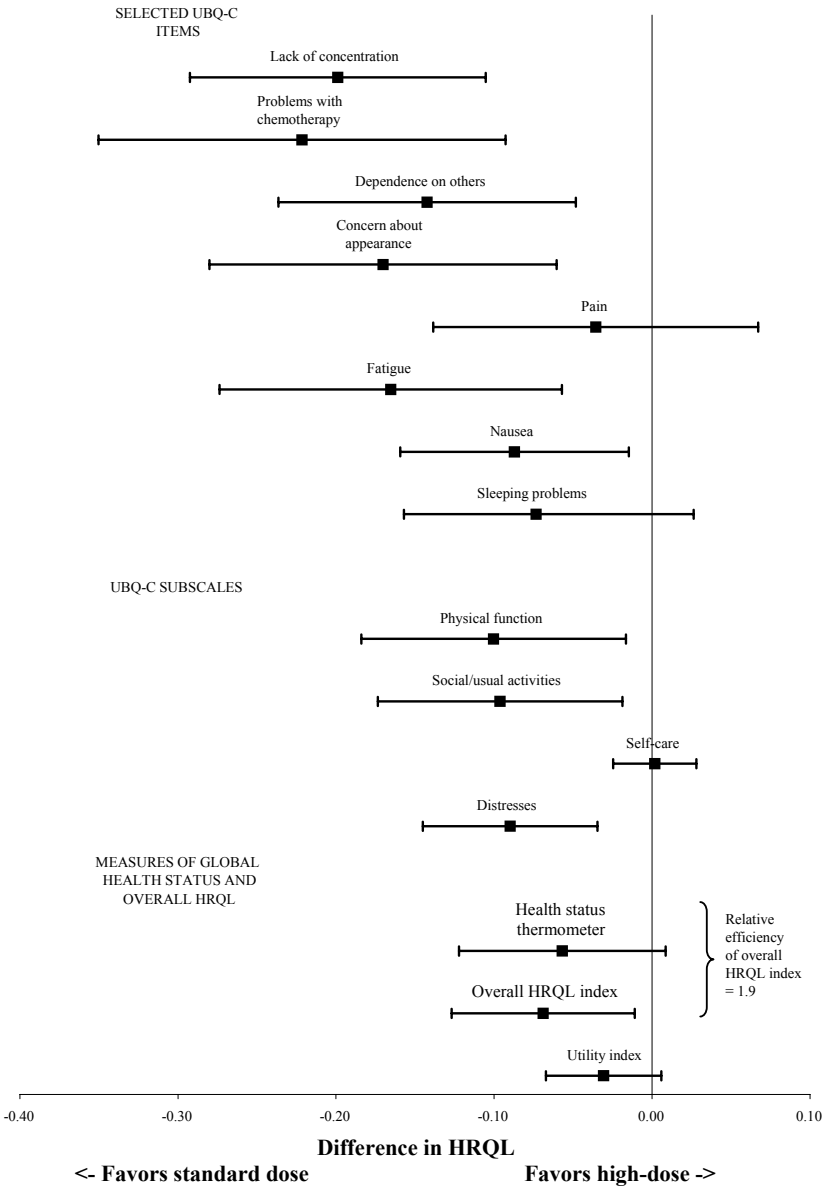


osMonths, survival in months. Good HRQL, score on utility index < 0.92. Poor HRQL, score on utility index ≥ 0.92.

6.4.4 Treatment comparison

The scoring algorithm was applied to the treatment comparison of high-dose chemotherapy versus standard-dose chemotherapy for early-stage breast cancer. Subjects receiving high-dose chemotherapy reported worse impairment of most specific aspects of HRQL (figure *6.2*), which was expected because high-dose chemotherapy is more toxic in this setting [136]. There was a trend to better mean scores on the utility index for patients allocated to standard-dose chemotherapy (mean 0.95) compared to high-dose chemotherapy (mean 0.92) with mean difference of -0.03 (95% CI -0.07 to 0.01, p = 0.10). The overall HRQL index gave stronger evidence of this effect (mean difference -0.07, 95% CI -0.13 to -0.01, t=2.36, p = 0.02) than the health status thermometer (mean difference -0.06, 95% CI -0.12 to 0.01, t=1.72, p = 0.09) (figure *6.2*). The relative efficiency of the overall HRQL index compared with the health status thermometer was 1.9. In this practical illustration, the improvement in precision by using the overall HRQL index compared to the health status thermometer was sufficient to conclude that the more toxic regimen causes significantly worse effects on overall HRQL.

Figure 6.2 Early cancer trial: differences in HRQL between treatment groups, based on: (i) UBQ-C items, (ii) UBQ-C subscales, (iii) health status thermometer, (iv) overall HRQL index, (v) utility index



HRQL, Health-related quality of life. All ratings on scale from 0 to 1.

6.5 Discussion

We have applied a scoring algorithm for a cancer-specific utility-based instrument to clinical trial datasets and illustrated how it can be optimised in different clinical contexts. The algorithm converts ratings from a cancer-specific questionnaire for HRQL into a utility index that is based on the perspective of cancer patients. First, we optimised the scoring algorithm in two different clinical contexts for breast cancer by adjusting the index weights using data from two clinical trials. Second, we applied the algorithm to generate utility scores. Third, we showed that the utility index had convergent validity with related scales from other instruments, discriminative validity between participants with differing performance status, responsiveness to toxic effects of chemotherapy in early cancer, and predictive validity about subsequent survival duration. Fourth, we used the utility index to inform a treatment comparison of high-dose chemotherapy with stem cell support versus standard-dose chemotherapy for high-risk early-stage breast cancer. It can be used to generate sensitive and responsive utility scores, and quality-adjusted life-years, that can be used within a trial to compare the net benefit of treatments and inform clinical decision-making.

The novelty of the approach described in this paper is that the scoring algorithm can be optimised for different clinical contexts. In contrast, most scoring algorithms for utility-based instruments use the same scoring algorithm across different diseases and treatments [9, 16] [103, 106-107]. The algorithm is optimised by giving additional weight to the subscales about specific aspects of HRQL that are most closely associated with a single-item global scale (the health status thermometer) in the relevant dataset. The reason to optimise the algorithm in different contexts is to reflect variations in patients' attitudes, preferences and priorities across different cancer types, stages, and treatments [32, 63].

The primary benefit of optimising the scoring algorithm for each clinical context is that it should better reflect the perspective of the individuals in that situation. For example, in the comparison of high-dose versus standard-dose chemotherapy for early-stage breast cancer (figure *6.2*), there were large differences in distresses and physical function but little or no difference in self-care. Combinations of the

subscales giving greater weight to self-care would yield little difference between high-dose and standard-dose chemotherapy, whereas those giving greater weight to distresses and physical function would favour standard-dose chemotherapy. We assigned weight according to correlations with the health status thermometer, resulting in significant differences between treatments on the indices for overall HRQL and utility which should reflect the preferences and attitudes of the women in the trial.

Optimising the scoring algorithm could give more precise estimates of clinically important differences in utility between patient groups, because the index is focussed on those aspects of HRQL that are most relevant to those patients. A more precise utility index will reduce the uncertainty around the incremental effectiveness of treatments in sensitivity analyses, because it is more responsive to small but meaningful effects of cancer treatments [172]. A more precise utility index will also reduce the sample size required to detect a given difference with a given level of precision [172].

Another benefit of optimising the scoring algorithm for each clinical context is that the ordering of the weights can inform clinicians and researchers about the importance of various symptoms, side effects and dysfunctions that patients in different clinical contexts most wish to avoid. For example, in both datasets we found that greatest weight was given to distresses, followed by social/usual activities, physical function and self-care (table *6.3*). The ordering of the weights assigned to each subscale may be related to several factors. The large weight assigned to distresses may reflect the emotional distress that most patients experienced due to having cancer, the physical symptoms of advanced cancer, and the side effects of toxic chemotherapy for early-stage cancer. The low weight assigned to self-care may reflect the lack of problems with self-care that most patients reported in each trial. The ordering of the weights may also reflect the number of items within each subscale, with distresses (21 items) assigned greater weight than physical function (3 items), social/usual activities (4 items) or self-care (1 item). There were some differences in weights between datasets. Greater weight was assigned to distresses and less weight was assigned to physical function for early cancer compared with advanced cancer (table 3). The greater weight assigned to distresses for patients with

early cancer may reflect their greater emotional distress due to a recent diagnosis of cancer, and their experience of side effects from chemotherapy which had not yet been administered to the patients with advanced cancer. The lower weight assigned to physical function for patients with early cancer is probably explained by the absence of the deterioration in physical function that occurs with advanced cancer. This information can be used by researchers to design more targeted interventions to improve HRQL in the dimensions of greatest importance to patients, and by all health care workers to improve counselling of patients [63].

We recommend that the scoring algorithm is optimised for each clinical context in which it is used. This is a potential limitation in that it requires additional analyses, and familiarity with Lumley's method. Another limitation is that the utility scores may not be comparable from one disease or treatment context to another, because the scoring algorithm and its index weights cannot be standardised across trials [32]. Consequently we recommend that our approach is used to compare treatments in the context of a trial in a well-defined population for a specific clinical condition, because the attitudes of patients are likely to be more similar. It is less suited to studies that include diverse populations, or for comparing utilities and quality-adjusted life-years from one study or context to another, because the attitudes of patients will be more diverse. Comparability of utility scores is a key requirement when utilities are used to inform economic decisions, because health funders and policy makers make decisions across diseases and contexts [9, 108]. However comparability of utility scores is less important when utilities are used to inform clinical decisions, because a clinical decision is always limited to a single disease type and stage. The requirements of utility scores used to inform clinical decisions are that they reflect the experiences of the patients under study, and are valid, sensitive and reliable.

The measurement properties of the utility index reported in this paper support its validity as a measure of HRQL for the clinical context of chemotherapy for early and advanced breast cancer. The utility index had convergent validity with independent scales of general health and global quality of life, was able to discriminate patients with different stages of cancer, and was responsive to changes attributable to having chemotherapy. Another way to validate a utility index is to compare the scores

derived by the utility index with utilities elicited directly from the same patients with a time trade-off interview. This could be performed in future studies.

Future research is also needed to determine whether optimisation of the scoring algorithm for each context makes a meaningful difference to the utility scores and QALYs generated from the utility index, their sensitivity and responsiveness to detect differences between treatment groups, and most importantly to the outcome of clinical decisions in specific clinical contexts.

Finally, it is important to comment on the strengths and limitations of the datasets used in this study. Patients participating in a clinical trial of treatments are the ideal source of information about the effects of those treatments on HRQL. The datasets included patients with early and advanced cancer, before, during and after chemotherapy. Compliance was good with questionnaire completion, particularly for the advanced cancer trial. We used validated cancer-specific questionnaires that included a broad range of items about specific aspects of HRQL that are commonly affected by cancer and side-effects of treatment. A limitation of the datasets is that they only included women in Australia and New Zealand with breast cancer receiving chemotherapy, so the results may not be applicable to other cancer types or treatments, other countries and men. Further application and validation of the utility index is ongoing in other clinical contexts including chemotherapy for advanced colorectal cancer and hormonal therapy for the prevention of breast cancer [126, 150]. The colorectal study includes British and male subjects. Compliance with completing questionnaires in the early cancer trial was poor during chemotherapy. Patients who do not complete questionnaires tend to have worse HRQL [137], so the analyses may underestimate the detrimental effects of treatment on HRQL. Finally, the early cancer trial is relatively old so the effects of chemotherapy may be different to that with more modern treatments. However the primary purpose of generating utility scores in this study was to illustrate the approach, rather than to inform clinical or health policy decisions.

Our approach enables HRQL data obtained with a simple questionnaire to be converted into utility scores by using an optimised scoring algorithm that reflects the perspective of the cancer patients under study. The approach is flexible and

applicable to other trials and other HRQL instruments. Generation of utility scores based on HRQL data collected within a clinical trial provides an ideal source of information to inform clinical decisions, and to add a useful additional perspective to inform health policy and economic decisions.

7. Comparing treatments in a randomised trial

7.1 Overview

This chapter further develops the scoring algorithm that was derived in chapter 5, applies it to data from a randomised controlled trial of chemotherapy for advanced breast cancer, and uses the results to evaluate the differences in overall HRQL and utility between subjects allocated to each treatment group in the trial. The analyses required two extensions to the previously described scoring algorithm. First, additional information about HRQL relevant to the treatment comparison was incorporated in the indices. Second, a method for deriving index weights for the scoring algorithm from longitudinal data was developed. The approach developed in this thesis was shown to be feasible for use in clinical trial evaluation, and adaptable to the incorporation of additional information about HRQL. The index gave a more precise estimate of differences between groups compared to a single-item global scale. Trends in scores for overall HRQL during treatment favoured the oral chemotherapy regimen tested in the trial, but the differences were not statistically significant. The scores on the utility index are being used to evaluate quality-adjusted time to progression, which is the primary endpoint for the trial.

7.2 Introduction

The previous three chapters described the development and validation of a scoring algorithm that converts responses to a cancer-specific HRQL questionnaire called the Utility-Based Questionnaire-Cancer (UBQ-C) into valid and precise indices of overall HRQL and utility. The indices are optimally weighted to reflect the relative importance of differing aspects of HRQL to patients, and express the desirability of cancer health states from the perspective of cancer patients. The scoring algorithm facilitates the evaluation and comparison of the net effect of cancer treatments tested in clinical trials in terms of overall HRQL and utility, and in terms of quality-adjusted survival by combining utilities and survival data.

The scoring algorithm developed in chapter 5 needed further development before it could be applied to the evaluation of a clinical trial. One limitation of the current scoring algorithm is that the indices of overall HRQL and utility may not incorporate relevant and potentially important information about particular aspects of HRQL that are likely to differ between specific study treatments. For example, it would be desirable to incorporate information about the convenience and acceptability of tablets versus injections in a trial of oral versus intravenous chemotherapy for advanced cancer. Another limitation of the current scoring algorithm is that the regression methods used to derive weights for the indices rely on the assumption that all the observations are independent of one another. It is valid to apply these regression methods to cross-sectional data (as in chapters 5 and 6), because the observations come from different patients so are independent. However the regression methods may give biased results when applied to longitudinal data that is obtained during a clinical trial, because successive measurements within a given patient are correlated but not independent.

The primary aim of the work reported in this chapter was to further develop and apply the scoring algorithm derived in chapter 5, to facilitate evaluation of a randomised controlled trial of treatments for advanced breast cancer that was described in chapter 3 (section 3.5.2). The scoring algorithm was made more relevant for the specific treatment comparison by incorporating additional information about the expected side effects of the treatments in the indices of overall HRQL and utility.

The scoring algorithm was extended so that it could be applied to longitudinal datasets obtained during clinical trials. The secondary aim of this chapter was to compare the scores for overall HRQL and utility in each treatment group during treatment. This comparison would determine which treatment is preferable in terms of overall HRQL, and provide utility values for the planned evaluation of the trial using quality-adjusted survival analyses.

7.3 Methods

7.3.1 Clinical trial design

The specific clinical trial evaluated in this chapter was the advanced cancer trial that was described in chapter 3 (section 3.5.2), and for which baseline data was analysed in chapter 6. To recap, this was a randomised controlled trial conducted by the Australia New Zealand Breast Cancer Trials Group (ANZBCTG) that compared two types of chemotherapy for women with advanced breast cancer. Eligibility criteria were reported in chapter 3 (section 3.5.2).

The intervention group received chemotherapy with capecitabine, which is a newer drug that has the advantage of oral administration. Patients take 2000mg/m^2 (about 6 to 8 tablets) administered orally for the first 14 days of each 21-day treatment cycle, or 1300 mg/m^2 (about 4 to 6 tablets) administered for all 21 days of each 21-day treatment cycle. Side effects are typically mild or moderate, but severe toxicity occurs in about 10% of patients.

The control group received chemotherapy with 'CMF', which is a traditional combination of intravenous and oral drugs that has been in use since the 1970s. The use of CMF has declined over the last ten years, as its administration is more complex than competing newer regimens and requires intravenous administration. Patients receive intravenous chemotherapy with methotrexate 40 mg/m^2 and 5-fluouracil 600 mg/m^2 on days 1 and 8, and oral cyclophosphamide 100mg/m^2 (about 3 to 4 tablets) for the first 14 days, of each 28-day treatment cycle. Side effects are different to capecitabine. They are also usually mild to moderate in severity, but severe in about 10% of patients.

Treatment was continued until disease progression, unacceptable toxicity, or intolerance. Time to progression was measured by clinical assessments every three to four weeks and imaging (computed tomography and bone scan) every 12 weeks. The schedule for assessment of HRQL is outlined in the next section.

The aim of the trial was to determine which chemotherapy regimen was preferable. It was hypothesised that capecitabine would be superior to CMF because of equivalent

or better tumour control, more convenient oral administration, and less troublesome side effects of treatment. The primary objective of the trial was to compare capecitabine with CMF in terms of quality-adjusted time to progression. This summary measure integrates improvements in HRQL due to relief of cancer symptoms, detriments in HRQL due to the side effects and inconvenience of treatment, and improvements in tumour control. It is potentially helpful as a measure of the net benefit of treatment that could assist patients and clinicians deciding between treatment options. Quality-adjusted time to progression is expressed in quality-adjusted life-years. It is calculated by applying standard quality-adjusted survival methods to HRQL data (expressed as a utility) and time to progression data (expressed in years) [75, 173]. The need to obtain utility data for this trial motivated the development and application of the scoring algorithm described in this thesis.

7.3.2 HRQL assessment

Subjects completed HRQL questionnaires prior to randomisation, then every three to four weeks until disease progression. This was done on day one of each treatment cycle (or up to seven days before) for subjects who were receiving study chemotherapy, and every four weeks for subjects who were no longer receiving study chemotherapy. HRQL questionnaires were not completed after disease progression. 'During treatment' was defined as from the day after randomisation until 30 days after disease progression.

The questionnaires that subjects completed were the UBQ-C, Chemotherapy Acceptability Questionnaire, and other questionnaires that were described in detail in chapter 3 (sections 3.2 and 3.3) and replicated in appendices 1 and 2. Only ratings on the UBQ-C and Chemotherapy Acceptability Questionnaire were used for analyses in this chapter. To recap, the UBQ-C is a validated, disease-specific HRQL questionnaire that includes 29 items grouped into four multi-item subscales about specific aspects of HRQL, and a global scale called the health status thermometer which is a single item that asks respondents for a unified assessment of their health status. The scores for the subscales are the simple average of the items, linearly transformed to a scale from 0 (worst) to 1 (best). The Chemotherapy Acceptability Questionnaire is a study-specific subscale that was designed to supplement the UBQ-C for this trial. It includes 15 items about the inconvenience and additional specific

side effects that were not assessed by existing instruments but were expected to occur with capecitabine or CMF. The score for the Chemotherapy Acceptability Questionnaire subscale, which will be referred to as 'chemotherapy acceptability', is the simple average of the items, linearly transformed to a scale from 0 (worst) to 1 (best).

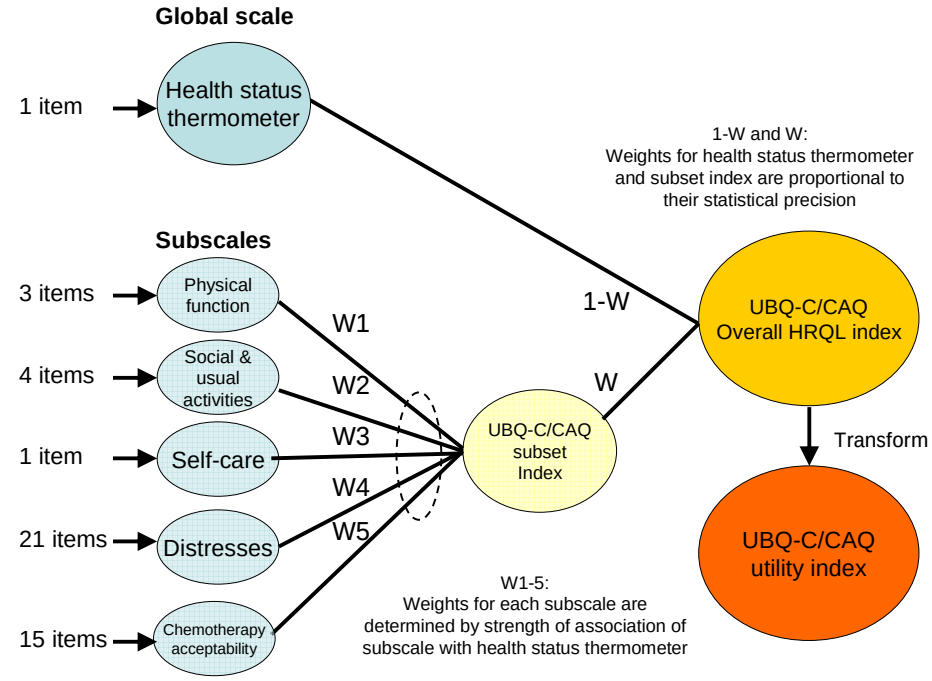
7.3.3 Further development of the scoring algorithm

This section reports how the scoring algorithm based on Lumley's approach [32] and derived in chapter 5 was extended to incorporate additional aspects of HRQL, and to be applied to longitudinal data.

Extending the scoring algorithm

The scoring algorithm was extended by incorporating the Chemotherapy Acceptability Questionnaire subscale, because it contributed additional information about aspects of HRQL that were expected to differ between treatment groups. The resultant indices were derived from a weighted combination of the health status thermometer, the four multi-item subscales of the UBQ-C, and the chemotherapy acceptability subscale (figure 7.1). The indices are referred to as the 'UBQ-C/CAQ subset index', 'UBQ-C/CAQ overall HRQL index' and 'UBQ-C/CAQ utility index' in the tables and figure of this chapter.

Figure 7.1 Deriving a utility index for the UBQ-C and CAQ questionnaires



UBQ-C, Utility-Based Questionnaire-Cancer. CAQ, Chemotherapy Acceptability Questionnaire.

Deriving subset index weights for longitudinal data

A subset index was derived from a weighted combination of the subscales of the UBQ-C and Chemotherapy Acceptability Questionnaire. Optimal weights (W1-5) were derived using ratings on the relevant subscales and the health status thermometer from the advanced cancer trial (figure 7.1). Analogous to the approach presented in chapters 5 and 6, the weights for the subscales of the UBQ-C and Chemotherapy Acceptability Questionnaire (W1 to W5 in figure 7.1) are designed to reflect the relative contribution of each subscale to overall HRQL, and were derived by giving additional weight to subscales that are more closely associated with the health status thermometer.

For the cross-sectional data analysed in chapters 5 and 6, there was one set of observations for each individual, and the weights reflected differences in the associations between people at one point in time. For the longitudinal data analysed in this chapter, there were multiple sets of observations for each individual, at multiple time points; so the weights should reflect the differences in the associations both between people and within each person over time.

Weights for cross-sectional data can be determined with simple multivariable linear regression methods, because the observations in different patients are independent. Weights for longitudinal data cannot be determined with simple multivariable linear regression, because successive observations within a given individual are correlated, not independent. Specifically, the observation on one occasion will give information about the observations on subsequent occasions. To analyse a series of such observations with simple linear regression as if they were independent may bias the magnitude and variance of the regression coefficients used to derive the weights [138].

There is no standard method to derive weights for longitudinal data using Lumley's approach [32], so the weights were derived using three different methods with varying levels of complexity: an 'optimal method', an 'uncorrelated method' and a 'baseline method'. The optimal method that will be described is valid for the analysis of longitudinal data because it avoids the problem of correlations between successive

measurements on a given individual. Because the optimal method is complex, in sensitivity analyses the weights were derived with two alternate methods: a simpler uncorrelated method that ignored the links between successive observations on a patient; and a baseline method that only used the cross-sectional data obtained at baseline. If the simpler methods gave similar weights, then they may be an appropriate alternative for future applications.

The optimal method is an adaptation of validated methods outlined by Sheppard and colleagues for the analysis of longitudinal data in an epidemiological context [174-175]. It avoids the problem of the correlation between successive measurements on a given individual by partitioning the association between the subscales and the health status thermometer into two components: First, the extent to which the differences *between people* in their subscale scores are reflected in the differences *between people* in their health status thermometer scores (*as in chapter 5 and 6*). Second, the extent to which the changes over time *within each person* in their subscale scores are reflected in the changes over time *within each person* in the health status thermometer scores (*not in chapter 5 or 6*).

The between-people and within-person associations were partitioned using Sheppard et al's approach. For each subscale, if we express the subscale score for the i^{th} person at the t^{th} timepoint as X_{it}, then the average personal subscale score for the i^{th} person over all timepoints is $\overline{X_i}$, and the within-person deviation from the average personal subscale score over time for the i^{th} person at the t^{th} timepoint is $\left(X_{it} - \overline{X_i}\right)$. $\overline{X_i}$ is used to determine the associations between people, and $\left(X_{it} - \overline{X_i}\right)$ is used to determine the associations within each person over time.

Multivariable ordinary least squares regression was performed. The dependent variable was the health status thermometer. The ten independent variables were $\overline{X_i}$ for each subscale, and $\left(X_{it} - \overline{X_i}\right)$ for each subscale. The between-people weight for each subscale was proportional to the regression coefficient of $\overline{X_i}$ for the corresponding subscale. The within-person weight for each subscale was proportional to the regression coefficient of $\left(X_{it} - \overline{X_i}\right)$ for the corresponding

subscale. The weight (W1 to W5) for each subscale was proportional to the sum of the within-person weight and between-person weight for the corresponding subscale. The subset index scores were then calculated by applying these weights to the subscale scores as follows:

[1] Subset index = [W1 * PF] + [W2 * SA] + [W3 * SC] + [W4 * DI] + [W5 * CAQ]

W1-5 are the weights for the subscales, PF is physical function, SA is social/usual activities, SC is self-care, DI is distresses, CAQ is chemotherapy acceptability.

In sensitivity analysis, the subset index weights that were derived with the optimal method were compared to weights derived with the two alternate methods that are simpler but potentially biased. The first was the uncorrelated method which ignored the link between successive measurements on a given individual. Multivariable ordinary least squares regression was performed using all data at baseline and during treatment. The dependent variable was the health status thermometer. The five independent variables were the five subscales. The weight for each subscale (W1-5) was proportional to the regression coefficient for the corresponding subscale. The advantage of the 'uncorrelated' method is that it is much simpler. The disadvantage is that the weights may be incorrect because the magnitude and variance of the regression coefficients may be biased.

The second alternate method was the baseline method that only analysed the cross-sectional data that was collected at baseline. The method described in chapter 5 was applied. Data collected during treatment was not used for this analysis. The advantage of this approach is that it is simple and statistically valid. The disadvantage is that the weights may be incorrect if the associations within subjects over time differ from the associations between subjects at baseline. This could conceivably occur if side effects that occur during treatment but are absent at baseline but adversely affect overall HRQL.

Deriving overall HRQL index weights

An overall HRQL index was derived by combining the subset index with the health status thermometer. The weights and the index were calculated using the formula [2] below, and the weights were applied to calculate an overall HRQL index for each participant at each time point using formula [3] below, as described in chapter 5:

[2] $W = Var(T) * [1 - r(T)] / MSE(R)$.

[3] Overall HRQL index = [W * Subset index] + [(1 – W) * HST]

W is the weight allocated to the subset index, so 1 – W is the weight allocated to the health status thermometer. Var(T) is the variance of the health status thermometer in the dataset. r(T) is the intraclass correlation coefficient of the health status thermometer, and was calculated with test-retest data from a previous validation study [13]. MSE(R) is the mean square for error from the multivariable ordinary least squares regression referred to above.

Transforming the overall HRQL index to the utility index

The transformation function derived in chapter 5 was applied to calculate a utility index from the overall HRQL index for each participant at each time point, as follows:

[4] Utility index = $1 - (1\text{-overall HRQL index})^{2.03}$.

7.3.4 Comparing ratings between treatment groups

Baseline

The comparison of baseline HRQL ratings between treatment groups was important in this clinical trial. Because of random allocation, baseline ratings on the health status thermometer, subscales, and indices of overall HRQL and utility were expected to be similar in each treatment group. However any clinically important imbalances between treatment groups at baseline suggest that they are dissimilar, and could confound the interpretation of the results [176]. Differences were considered to be potentially clinically important when the absolute difference in mean scores was greater than 0.05 on a scale from 0 to 1, consistent with recommendations from others as to the magnitude of a clinically meaningful difference [176]. Because any

imbalances can be attributed to chance, it was not appropriate to compare ratings using statistical tests.

During treatment

To compare HRQL between treatment groups during treatment, the mean ratings during treatment of subjects allocated to capecitabine and CMF on the health status thermometer, subscales, and indices of overall HRQL and utility were compared using generalised estimating equations that took into account the correlation between the successive observations for each participant [138, 177]. Differences were considered to be statistically significant when the p-value was < 0.05. In the primary analysis, the mean ratings during treatment were adjusted for baseline data where it was available, because this accounts for chance imbalances in HRQL between treatment groups at baseline. Because there is uncertainty about the desirability of adjusting for baseline data in HRQL evaluation of clinical trials [178-179], the effect of not adjusting for baseline data was tested in a sensitivity analysis.

7.3.5 Evaluating the precision of the overall HRQL index

A feature of the scoring algorithm described in this thesis is that the overall HRQL index compared with the health status thermometer should give an estimate of the differences in mean scores between subjects in each treatment group that is more precise (narrower confidence intervals) but unbiased (similar point estimates). This hypothesis was tested in two related ways. First, the magnitudes of the standard errors of the health status thermometer and of the overall HRQL index for the differences between treatment groups during treatment were compared. A more precise scale will have a smaller standard error, and correspondingly narrower confidence intervals [138]. Second, the relative precision of each measure was compared using the relative efficiency statistic. As described in chapter 5, the relative efficiency statistic is the factor by which the sample size can be reduced when the more precise and efficient scale is used [138, 144]. The relative efficiency statistic was calculated as the squared ratio of the z-score of the overall HRQL index and the z-score of the health status thermometer for the difference between treatment groups during treatment.

7.4 Results

7.4.1 Study profile and patient characteristics

The study profile and patient characteristics were presented in detail in chapter 3 (sections 3.5.2 and 3.6). Of the 325 patients that were enrolled, 216 subjects were allocated to capecitabine, and 109 subjects were allocated to CMF. Two subjects were excluded from the primary analysis reported elsewhere [124], because one did not have breast cancer and one received a precluded study drug, but all subjects were included in the analyses reported in this chapter.

Compliance with completion of forms was excellent (figure 3.3 of chapter 3). The UBQ-C was completed at baseline (prior to randomisation) by 295 of 311 subjects who were expected to complete it (96%). The UBQ-C was completed 3227 times during treatment by 299 of 311 subjects who were expected to complete it (96%). 16 completed it once, 66 completed it two to four times, 95 completed it five to nine times, and 122 competed it ten or more times. Compliance with form completion was similar for subjects allocated to capecitabine or CMF.

7.4.2 Ratings on the UBQ-C and Chemotherapy Acceptability Questionnaire

Subjects' mean ratings on the UBQ-C and Chemotherapy Acceptability Questionnaire for each treatment group at baseline are shown in table 7.1. Mean scores at baseline were similar between treatment groups for physical function, social/usual activities, and chemotherapy acceptability. This was expected because of randomisation. Mean scores at baseline were marginally better for subjects allocated to CMF than those allocated to capecitabine for the health status thermometer and distresses. The effect of this small chance imbalance on ratings during treatment was tested in the next section.

Table 7.1 Advanced cancer trial at baseline: ratings by treatment group

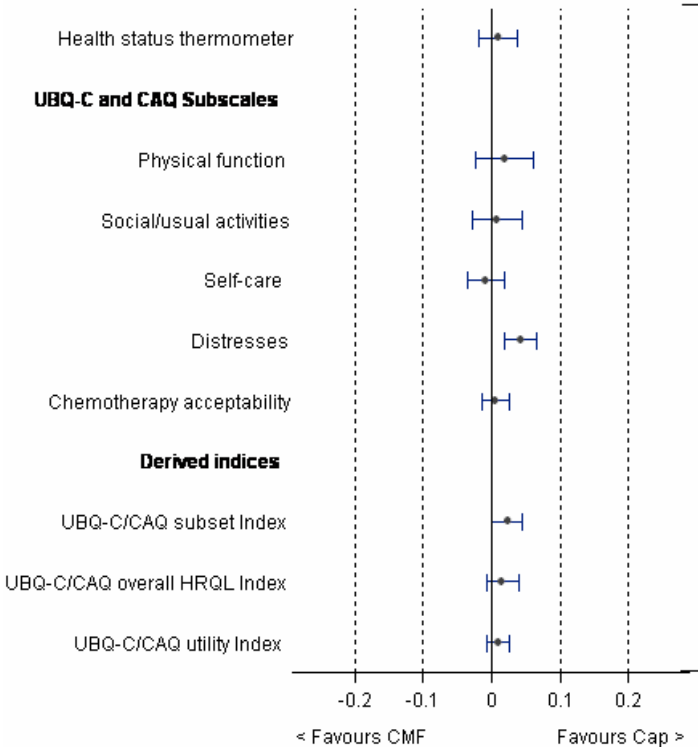
Treatment arm	Capecitabine		CMF			
	Mean	**SD**	**Mean**	**SD**	**Mean difference**	**95% CI**
Health status thermometer	0.68	0.20	0.71	0.18	0.03	-0.02, 0.07
UBQ-C, CAQ subscales						
Physical function	0.53	0.32	0.54	0.34	0.01	-0.07, 0.09
Social/usual activities	0.66	0.28	0.66	0.31	-0.01	-0.08, 0.06
Self-care	0.89	0.20	0.88	0.22	-0.01	-0.06, 0.04
Distresses	0.78	0.16	0.80	0.14	0.02	-0.02, 0.06
Chemotherapy acceptability	0.90	0.10	0.91	0.10	0.01	-0.02, 0.03
Derived indices						
UBQ-C/CAQ subset index	0.78	0.15	0.79	0.15	0.01	-0.03, 0.05
UBQ-C/CAQ overall HRQL index	0.72	0.17	0.74	0.16	0.02	-0.02, 0.06
UBQ-C/CAQ utility index	0.89	0.12	0.91	0.09	0.02	-0.01, 0.04

CMF, cyclophosphamide, methotrexate and 5-fluorouracil. SD, Standard deviation. 95% CI, 95% confidence intervals. UBQ-C, Utility-Based Questionnaire-Cancer. CAQ, Chemotherapy Acceptability Questionnaire. HRQL, heath-related quality of life. All ratings on scale from best (one) to worst (zero).

Subjects' mean ratings on the UBQ-C and Chemotherapy Acceptability Questionnaire for each treatment group during treatment, adjusted for baseline, are shown in table 7.2. Subjects in both treatment groups reported worst impairment for physical function, and least impairment for self-care compared with other subscales. Mean ratings were significantly better for capecitabine than CMF for the distresses subscale ($p=0.0003$). Mean ratings were similar for capecitabine and CMF for subscales about physical function, social and usual activities, and chemotherapy acceptability ($p > 0.2$).

The results of a sensitivity analysis, whereby subjects' mean ratings on the UBQ-C and Chemotherapy Acceptability Questionnaire for each treatment group during treatment were not adjusted for baseline values, are shown in table 7.3. The unadjusted mean ratings on each subscale and index were generally about 0.04 lower than the adjusted mean ratings. The point estimate of the differences in mean ratings between treatment groups on each subscale and index was shifted by about 0.02 in favour of CMF. As a result, the differences for distresses were no longer statistically significant ($p=0.12$), and the trend to better mean ratings on capecitabine for other subscales were no longer apparent.

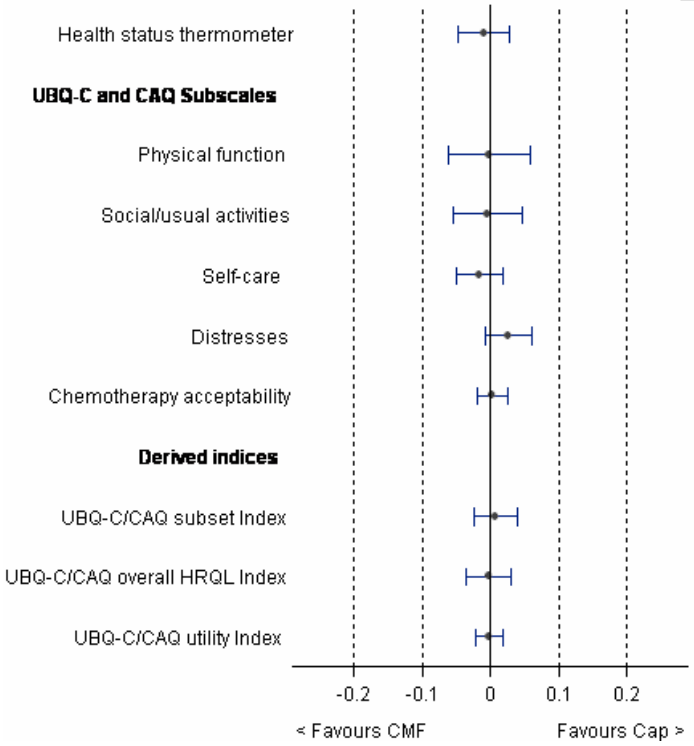
Table 7.2 Advanced cancer trial during treatment: ratings by treatment group (adjusted for baseline)



Treatment arm	Capecitabine	CMF	Difference		
n	201	98			
	Mean	**Mean**	**Mean**	**Standard error**	**p-value**
Health status thermometer	0.737	0.728	0.009	0.015	0.5
UBQ-C and CAQ Subscales					
Physical function	0.570	0.551	0.020	0.022	0.4
Social/usual activities	0.710	0.703	0.008	0.019	0.7
Self-care	0.915	0.923	-0.009	0.013	0.5
Distresses	0.852	0.809	0.043	0.012	0.0003
Chemotherapy Acceptability	0.890	0.884	0.006	0.010	0.6
Derived indices					
UBQ-C/CAQ subset index	0.820	0.797	0.023	0.011	0.04
UBQ-C/CAQ overall HRQL index	0.770	0.755	0.015	0.012	0.20
UBQ-C/CAQ utility index	0.927	0.918	0.010	0.008	0.22

Figure shows mean difference between treatment groups and 95% confidence intervals. Abbreviations overleaf.

Table 7.3 Advanced cancer trial during treatment: ratings by treatment group (without adjustment for baseline)



Treatment arm	Capecitabine	CMF	Difference		
n	201	98			
	Mean	**Mean**	**Mean**	**Standard error**	**p-value**
Health status thermometer	0.714	0.724	-0.010	0.019	0.6
UBQ-C and CAQ Subscales					
Physical function	0.536	0.538	-0.002	0.031	0.9
Social/usual activities	0.675	0.679	-0.004	0.026	0.9
Self-care	0.898	0.914	-0.016	0.017	0.3
Distresses	0.831	0.804	0.026	0.017	0.12
Chemotherapy Acceptability	0.881	0.878	0.003	0.011	0.8
Derived indices					
UBQ-C/CAQ subset index	0.796	0.788	0.007	0.016	0.6
UBQ-C/CAQ overall HRQL index	0.746	0.750	-0.003	0.017	0.8
UBQ-C/CAQ utility index	0.913	0.915	-0.003	0.010	0.8

Figure shows mean difference between treatment groups and 95% confidence intervals. Abbreviations overleaf.

Legend for tables 7.2 and 7.3

CMF, cyclophosphamide, methotrexate and 5-fluorouracil. Cap, capecitabine. UBQ-C, Utility-Based Questionnaire-Cancer. CAQ, Chemotherapy Acceptability Questionnaire. HRQL, heath-related quality of life. All ratings on scale from best (one) to worst (zero).

7.4.3 Weights for health status thermometer, subscales and subset index

The weights assigned to the health status thermometer, subscales and subset index with the optimal, uncorrelated and baseline methods are shown in table 7.4. For the optimal method, greatest weight was given to the distresses subscale, and least weight was given to the self-care subscale. Therefore distresses were most strongly associated with health status, and self-care was least strongly associated with health status. The between-people weights (sum 56%) were slightly greater than the within-person weights (sum 44%). Therefore the differences in subscale ratings between people explained slightly more of the variation in health status than the differences in subscale ratings within each individual over time. The weights derived with the 'optimal' and 'uncorrelated' methods were almost identical. The 'baseline' method assigned slightly higher weights to the health status thermometer and subscale for physical function, and lower weights to self-care and distresses. This means that self-care and distresses were more strongly associated with the health status thermometer during the treatment phase than at baseline.

Table 7.4 Advanced cancer trial during treatment: weights for scoring algorithm

		Optimal method			Uncorrelated method	Baseline method
		Between-people weights	*Within-person weights*	Weight	Weight	Weight
W1	Physical function	*0.06*	*0.03*	0.10	0.11	0.18
W2	Social/usual activities	*0.08*	*0.06*	0.15	0.15	0.15
W3	Self-care	*0.04*	*0.02*	0.07	0.07	0.02
W4	Distresses	*0.18*	*0.24*	0.42	0.44	0.39
W5	Chemotherapy acceptability	*0.19*	*0.08*	0.27	0.24	0.26
W	Health status thermometer			0.61	0.61	0.66
1-W	Subset Index			0.39	0.39	0.34

1-W, W, W1-5 refer to the weights assigned to the health status thermometer, subset index, and subscales in formulae [1] and [2] (see text of section 7.3.3, and figure 7.1)

7.4.4 Treatment comparison

Subjects' mean scores on the overall HRQL index and utility index for each treatment group at baseline were shown in table 7.1 above. Mean scores at baseline were marginally better for both indices amongst subjects allocated to CMF than those allocated to capecitabine. The effect of this small chance imbalance on ratings during treatment was tested in the next section.

Mean scores on the overall HRQL index and utility index for each treatment group during treatment, adjusted for baseline, were shown in table 7.2 above. There was only a weak trend to better ratings for those allocated capecitabine than those allocated CMF for the overall HRQL index ($p = 0.2$) and the utility index ($p=0.22$). The mean rating for the utility index for those allocated capecitabine was 0.93 (95% CI 0.92 to 0.94), and for those allocated CMF was 0.92 (95% CI 0.91 to 0.93).

The results of a sensitivity analysis, where subjects' scores during treatment were not adjusted for baseline values, were shown in table 7.3 above. The point estimate of the differences between treatment groups on each index were shifted slightly in favour of CMF. As a result, the trend to better mean scores on capecitabine for the overall HRQL index was no longer apparent.

7.4.5 Comparing precisions of the overall HRQL index and health status thermometer

The overall HRQL index estimated the difference in mean scores more precisely than the health status thermometer by two measures. First, the standard error was smaller: 0.012 versus 0.015. This is illustrated by the narrower confidence intervals for the overall HRQL index than the health status thermometer in tables 7.2 and 7.3. Second, the relative efficiency statistic derived from the z-scores for the health status thermometer (of 0.65) and the overall HRQL index (of 1.28) was 3.9. As outlined in chapter 5, a relative efficiency statistic of 3.9 corresponds to a 74% reduction in the sample size needed to detect a significant difference between treatment groups by using the overall HRQL index, rather than the health status thermometer.

7.5 Discussion

In this chapter, the scoring algorithm that was developed in chapter 5 was extended by incorporating relevant information about the acceptability and convenience of treatment to subjects in a randomised trial comparing oral and intravenous chemotherapy for advanced breast cancer, and further developed so that it could be applied to the longitudinal data obtained during the trial. Average overall HRQL and utility of subjects allocated to the two chemotherapy regimens was compared. Differences between treatment groups were found in some aspects of HRQL but not others, which were reflected in the small differences between treatment groups for the indices of overall HRQL and utility. This was consistent with other global measures that found no statistically significant difference between treatment groups [124]. Consistent with findings in chapter 5, the derived overall HRQL index compared to the single-item health status thermometer gave an estimate of differences between treatment groups that was more precise but unbiased.

The key development of the scoring algorithm in this chapter was the derivation of a method to derive optimal index weights from longitudinal data. The indices of overall HRQL and utility give additional weight to subscales that are more closely associated with the health status thermometer. They reflect differences between people at baseline, differences between people during treatment, and changes within each person over time. It can be argued that the associations within each person over time are the most relevant and important, because this variation reflects the effects of cancer and its treatment on the HRQL of each patient. It was uncertain if the associations within each person over time would be similar or different to the associations between people at baseline and between people during treatment. Application of the optimal method demonstrated that each of these associations were quite similar. The uncorrelated method of deriving index weights is not statistically valid, but is simpler and gave similar weights. If the weights are similar in future applications then the uncorrelated approach could be recommended as the standard approach in preference to the more complex optimal method. The baseline method gave slightly different weights, which implies that the aspects of HRQL that were most important to health status were not the same at baseline as during treatment.

Therefore we recommend at this time that weights should be derived using the optimal method from all data that will be analysed.

A key extension of the scoring algorithm in this chapter was to incorporate additional information that was relevant for a specific treatment comparison. The indices of overall HRQL and utility incorporated a subscale for chemotherapy acceptability, in addition to the UBQ-C subscales for physical function, social and usual activities, self-care and distresses. Information about the convenience and acceptability of chemotherapy was relevant for the trial, which compared oral versus intravenous chemotherapy, and was also found to be important as reflected by the weight given to it.

Ratings for distresses were significantly better for capecitabine than CMF. The distresses subscale is comprised of symptoms relating to the side effects of treatment and the symptoms of metastatic cancer. Better ratings for distresses indicate that on balance these symptoms were better controlled on capecitabine. Ratings for physical function, social and usual activities, and self-care were similar for each treatment group, which indicates that these aspects were equally affected on each regimen. Ratings for chemotherapy acceptability were also similar for each treatment group. This was an unexpected finding, because capecitabine has more convenient oral administration and was expected to be more acceptable to patients. The lack of difference suggests that the convenience of oral administration is less important than anticipated.

There was a reasonable expectation that overall HRQL and utility would differ between treatments because of their different side effect profile, methods of administration and anti-cancer effects. This expectation was supported by better ratings for distresses for capecitabine, but was not supported by ratings for the health status thermometer, which were similar in each treatment group. It was hypothesised that the lack of a significant difference was due to the imprecision of the health status thermometer. The analyses found a trend to better overall HRQL on capecitabine, which was not statistically significant (mean difference 0.015 on a scale from 0 to 1, $p=0.2$). This implies that the side effects and inconvenience of therapy, and the

control of cancer-related symptoms, were similar or better for capecitabine versus CMF, but that the differences were not beyond the play of chance.

The mean scores on the utility index in this study were 0.92 for the CMF group and 0.93 for the capecitabine group. These scores equate to the subjects in the trial, who were suffering from metastatic breast cancer, being willing to trade-off less than 10% of their life-expectancy in order to return to full health. The validity of these scores has not been validated by direct comparisons with time trade-off interviews with trial subjects, but have face validity in that they are consistent with utilities reported in the few other studies that were valued by patients with metastatic breast cancer using direct utility interviews (table 7.5). Perez repeatedly interviewed 38 New Zealand patients with metastatic breast cancer over a 12 month period, using a time trade-off interview with an unusual 30 day horizon, and found a similar mean utility over time of 0.95 for those receiving chemotherapy [180]. Lidgren interviewed 61 Swedish patients with metastatic breast cancer using a time trade-off interview with 10 year horizon, and obtained a mean utility of 0.78 for those on chemotherapy [181].

The mean scores on the utility index in this study were higher than utilities for metastatic breast cancer reported in studies when assigned by lay people using a direct utility-based scaling method, or by a utility-based instrument valued from the lay people perspective (table 7.5). The mean utilities in this study were also generally higher than utilities for metastatic breast cancer reported in studies when valued by experts by direct utility interview (table 7.5). This is expected because patients consistently assign higher utilities to cancer heath states than members of the general population and experts, as was described in chapter 2 (sections 2.4.3 and 2.5.4). Another explanation is the use of a time trade-off with a horizon of two years rather than five years, which tends to give higher values (section 2.4.2, table 7.5). A two-year horizon was used because it is consistent with the expected median survival of the group (section 3.4).

Table 7.5 Comparison of utilities for advanced breast cancer

Perspective Author, Year	Reference	Elicitation method	Comments	Specific to chemotherapy?	Utility value
Patient					
Perez 2001	[180]	TTO	30-day horizon	N	0.95
Grimison	*This paper*	UBQ-C		Y	0.90
Lidgren 2007	[181]	TTO	10-year horizon	Y	0.78
Lay people					
Lloyd 2006	[182]	SG		Y	0.72
Lidgren 2007	[181]	EQ-5D	By patients/UK lay people tariff	Y	0.69
Dranitsar 2000	[183-185]	TTO	By lay people & experts	Y	0.67
Grann 1999	[186-188]	TTO		N	0.52
Milne 2006	[189]	EQ-5D	By lay people	Y	0.51
Milne 2006	[189]	TTO		Y	0.49
Milne 2006	[189]	VAS		Y	0.46
van den Hout 2003	[190-192]	EQ-5D	By patients/UK lay people tariff	Y	0.40
Cykert 2004	[193]	SG		N	0.30
Expert					
Hillner, 2000	[194]	Estimate		N	1.00
Launois 1996	[195-196]	SG	By nurses	Y	0.75
Brown 1998	[196-201]	SG	By nurses	Y	0.70
Hillner 1991	[202-205]	Estimate		Y	0.70
Dranitsar 2000*	[183-185]	TTO	By lay people & experts	Y	0.67
Hutton 1996	[196, 206]	SG	By nurses	Y	0.62
Lonning 2006	[207]	Estimate		N	0.50

Perspective, perspective from which utility is elicited. Author, first author of reference study. Year, year of publication of reference study. 'Specific to chemotherapy?,' did study elicited values specifically for patients with metastatic breast cancer receiving chemotherapy? Y, yes. N, no. TTO, time trade-off. UBQ-C, Utility-Based Questionnaire-Cancer. SG, Standard gamble. VAS, visual analogue scale. EQ-5D, EuroQol EQ-5D. *, record listed under 'lay people' and 'expert' as utilities elicited from both.

The utility scores derived in this chapter will be used to inform important forthcoming trial analyses of quality-adjusted time to progression and quality-adjusted survival. There is a reasonable expectation that quality-adjusted PFS and quality-adjusted overall survival will be better on capecitabine compared to CMF, because utility, determined in this chapter, and PFS and overall survival, reported previously [124], are at least as good or better on capecitabine.

In summary, the scoring algorithm developed in this thesis has been shown to be feasible for the evaluation of a clinical trial, and the resultant utility scores have face validity. Ongoing work will use the results to inform quality-adjusted PFS and survival analyses of the trial. The next chapter summarises the work presented in this thesis, considers its strengths and limitations, and discusses its implications for practice and future research.

8. Discussion

8.1 Overview

This chapter begins in section 8.2 by revisiting the rationale for, aims of, and general approach taken to the work reported in the thesis. Section 8.3 summarises the principal findings. Section 8.4 considers the strengths and limitations of the general approach taken compared with alternatives. Section 8.5 postulates the likely implications of a utility-based instrument that is based on the perspective of cancer patients, optimally weighted for specific clinical contexts, and feasible for use in clinical trials. Section 8.6 identifies priorities for future research. The chapter ends with concluding remarks.

8.2 Revisiting the rationale for, aims of, and approach taken to the thesis

The background chapter (chapter 2) presented the rationale for using utilities to evaluate cancer treatments in terms of the potential trade-offs between benefits and harms. First, utilities provide a stand-alone assessment of the net effect of a treatment on overall health-related quality of life (HRQL) that can be used to quantitatively value relief of cancer symptoms and improvements in physical and psychological function, versus the impact of treatment-related side effects. Second, utilities can be combined with survival data to simultaneously value effects on survival and HRQL, in terms of quality-adjusted life years (QALYs). Third, utilities can be combined with survival and cost data to simultaneously value effects on survival, HRQL, and the costs of treatment, in terms of cost per additional QALY.

The background chapter also emphasised that the source of utilities used to evaluate cancer treatments is important, because variations can influence the outcome of decision-making in two important ways [208]. One important source of variation in utilities is the perspective from which they are valued. Patients generally assign a higher utility to a given health state than lay people. The utility assigned to a health state influences the magnitude of the incremental effectiveness and cost-effectiveness of a treatment in QALYs and cost per additional QALY. Another important source of variation in utilities is the approach taken to derive the utility. Disease-specific, utility-based instruments may give more precise estimates than generic utility-based instruments or direct utility-based scaling methods such as the standard gamble or time trade-off, if they are more sensitive to changes in health status due to the effects of that disease and its treatment. The precision of a utility estimate influences the uncertainty around estimates of incremental effectiveness and cost-effectiveness.

It was argued in chapter 2 that cancer treatments compared in clinical trials should be evaluated using utilities that are obtained from trial participants, because they have experienced the relevant health states. A utility measure in a clinical trial needs to be sufficiently responsive to detect small but meaningful changes in HRQL due to the effects of cancer treatments, and to be feasible for completion on a repeated basis by trial participants who are often unwell.

The work presented in this thesis was motivated by the need to determine the differences in utilities and QALYs between chemotherapy regimens being compared in clinical trials for early and advanced breast cancer that were described in chapter 3.

The general aim of the thesis, as presented in chapter 1 (section 1.2), was to develop a scoring algorithm that converts the responses to a cancer-specific questionnaire into an optimally-weighted utility index. The index was intended to:

i) reflect the perspective of patients with cancer;

ii) be optimally weighted for comparisons in a specific clinical context, *and*

iii) be feasible for use in cancer clinical trials.

The main approaches of the thesis, as outlined in chapter 4 (section 4.3), were to:

1. Select the items of the Utility-Based Questionnaire-Cancer (UBQ-C) from which to derive the utility index
2. Use data from a valuation survey of cancer patients to value the health states described by the UBQ-C
3. Produce a scoring algorithm that assigns a utility index score to health states described by the UBQ-C
4. Optimise the scoring algorithm in specific clinical contexts using data from randomised trials
5. Validate the utility index using related measures of HRQL and other factors
6. Apply the utility index to treatment comparisons using data from two randomised trials

The next section summarises the findings of the thesis in relation to the approach taken and general aims.

8.3 Summary of principal findings

This section summarises the principal findings of the thesis.

A utility index was derived from the health status thermometer, and the multi-item subscales for physical function, social/usual activities, self-care and distresses from the UBQ-C. Each subscale is the simple average of its one to 21 component items. The UBQ-C is a validated cancer-specific HRQL questionnaire that was designed to be feasible for use in clinical trials of cancer treatments.

204 patients with advanced cancer rated their current health status and HRQL by completing the UBQ-C, and assigning time trade-off utilities, in a valuation survey (chapter 3). The mean time trade-off utility was 0.91 and the median was 0.995. Half of the patients assigned a utility of 1.0 to their current health state. The valuation survey involved patients with cancer because the resultant utility index was intended to reflect the perspective of patients with cancer.

A scoring algorithm was derived that assigned scores on the utility index to health states described by the UBQ-C (chapter 5), using data from the valuation survey and adapting a methodological approach developed by Lumley et al [32]. First, a subset index was derived from a weighted combination of the UBQ-C subscales for physical function, social/usual activities, self-care and distresses. The weight assigned to each subscale was proportional to its correlation with the health status thermometer. Second, an overall HRQL index was derived from a weighted combination of the health status thermometer and the subset index. The weights assigned were proportional to the statistical precision of each measure. Third, a utility index was derived by applying a power transformation to the overall HRQL index. The formulae were as follows:

$$\text{Subset index} = [W1 * PF] + [W2 * SA] + [W3 * SC] + [W4 * DI]$$

$$\text{Overall HRQL index} = [\ W * \text{Subset index}\ ] + [\ (1 - W\ ) * HST\ ]$$

$$\text{Utility index} = 1 - (1 - \text{overall HRQL index})^{2.03}$$

W1-4 are the weights for the subscales, PF is physical function, SA is social/usual activities, SC is self-care, DI is distresses, HST is the health status thermometer. W is the weight allocated to the subset index, so 1 – W is the weight allocated to the health status thermometer.

Formulae to calculate W and 1-W were presented in chapter 5 (section 5.3.3).

In the valuation survey, the mean of the utility index scores was similar to the mean of the TTO utilities (0.92 versus 0.91, p=0.6). The mean absolute error was 0.10, and half of the predicted utilities were within 0.05 of assigned time trade-off utilities. The utility index was substantially correlated with other measures of HRQL.

The scoring algorithm for the utility index was optimised to reflect the attitudes of subjects in specific clinical contexts in chapter 6. The contexts were advanced breast cancer before chemotherapy; and early breast cancer before, during and after chemotherapy. The algorithm was optimised by adjusting the weights assigned to the subscales to reflect the correlations with the health status thermometer in each dataset. The weight assigned to the subset index and health status thermometer were also adjusted to reflect the precision of each measure in each dataset. Optimal weights for each trial are shown again in table 8.1.

Table 8.1 Included studies: comparison of weights for scoring algorithm

	Study	**Valuation survey**	**Early cancer trial**	**Advanced cancer trial**	**Advanced cancer trial**
	Clinical context	Advanced cancer (chapter 5)	Chemotherapy for early breast cancer (chapter 6)	Before chemotherapy for advanced breast cancer (chapter 6)	During chemotherapy for advanced breast cancer (chapter 7)
Weights for subset index					
W1	Physical function	0.28	0.20	0.09	0.10
W2	Social/usual activities	0.06	0.23	0.25	0.15
W3	Self-care	0.01	0.04	0.01	0.07
W4	Distresses	0.64	0.53	0.64	0.42
W5*	Chemotherapy acceptability	-	-	-	0.27
Weights for overall HRQL index					
1-W	Health status thermometer	0.65	0.66	0.63	0.61
W	Subset index	0.35	0.34	0.37	0.35

W1-5, 1-W, W refer to the weights assigned to the health status thermometer, subset index, and subscales in formulae in section 8.3. * Only included in one context.

The utility index was validated by application to the early and advanced cancer trials using related measures of HRQL and other factors in chapter 6. The utility index discriminated between breast cancer that was advanced rather than early (means 0.88 versus 0.94, $p < 0.0001$), and was responsive to toxic effects of chemotherapy in early breast cancer (mean change 0.07, $p<0.0001$). It also had convergent validity with related scales from other instruments, discriminative validity between participants with differing performance status, and predictive validity about subsequent survival duration.

The indices were applied to inform a treatment comparison of high-dose chemotherapy with stem cell support versus standard-dose chemotherapy for high-risk early-stage breast cancer in chapter 6. The utility index showed a trend towards better mean scores for standard-dose chemotherapy (means 0.95 versus 0.92, $p=0.1$). The indices were also applied to inform a treatment comparison of chemotherapy with oral capecitabine versus intravenous CMF for advanced breast cancer in chapter 7. This required two modifications to the scoring algorithm. The utility index was extended by incorporating additional information about the specific side effects of chemotherapy for advanced breast cancer, and the algorithm was optimised using regression methods that accounted for the correlations between repeated measures from individual subjects. Optimal weights are shown in the last column of table 8.1. The utility index showed no significant difference between treatment groups. The mean utility scores were 0.927 for capecitabine and 0.918 for CMF (mean difference 0.10, 95% CI -0.01 to 0.03, $p=0.22$).

8.4 Strengths and limitations of approach compared with alternatives

8.4.1 Determining the items and response options

The utility index was derived from the subscales of the UBQ-C, which is a cancer-specific HRQL questionnaire. The UBQ-C has several strengths as the basis for a utility-based instrument. It includes items about a broad range of aspects of HRQL that are relevant to cancer patients, as well as a single-item global scale of health status. The items have established evidence of feasibility, reliability and validity as reported in chapter 3 (section 3.3). It focuses on cancer-specific aspects of HRQL likely to be affected by cancer and its treatment, which should increase the ability of the utility index to detect small but meaningful differences in HRQL between treatment groups, and subtle changes over time, as discussed in chapter 2 (section 2.5.3). It has been shown to be feasible for use in clinical trials in other contexts including chemoprevention for breast cancer [126], and chemotherapy for testicular and colorectal cancer [123, 150]. Its use in these trials enables the use of data already collected to further establish the validity of the utility index, as was done in chapters 6 and 7, and to derive utilities and QALYs to inform decisions about cancer treatments.

An alternative approach would be to derive the utility index from one of the other cancer-specific HRQL instruments such as the EORTC QLQ-C30 or FACT-G that were described in chapter 2 (table 2.1). Because these instruments have been used more extensively, the scoring algorithm could be applied to data from a larger number of trials to generate utility scores for more clinical contexts. Another approach would be to derive the utility index from a generic instrument. This would improve the comparability of the scores across diseases, but would reduce the descriptive richness of the instrument and may limit the ability of the utility index to detect cancer-related effects.

The utility index was derived from the subscales of the UBQ-C, rather than from individual items as done by others who have adapted a profile-based HRQL instrument [95, 98, 107]. Producing a scoring algorithm from four subscales rather than from a larger number of individual items minimised problems with collinearity. One disadvantage of using subscales rather than items is that they provide less

information about the relative importance of different aspects of HRQL. For example, trade-offs between cancer symptoms and treatment side-effects that could be understood by examining the weight assigned to each individual item were concealed because those items are all contained within the distresses subscale.

8.4.2 The valuation survey

The data used to value the health states described by the UBQ-C, and to produce the scoring algorithm, were taken from a valuation survey of patients with cancer. These respondents are the ideal source of patient-based valuations about cancer health states because they have direct experience of those health states. The valuation survey was restricted to patients with advanced cancer, so the scoring algorithm may be less applicable to those with cancer that is early, or in remission, or those who are at increased risk of developing cancer. An alternative approach to increase generalisability would have been to obtain utilities from people in all these situations. The disadvantage of this approach is that the attitudes of people in different situations may differ. Another approach would have been to conduct a separate valuation survey and produce separate scoring algorithms in other contexts, which would have required substantial additional resources. Instead, the scoring algorithm was optimised in different contexts by adjusting the weights based on correlations with the health status thermometer.

The ratings assigned by subjects in the valuation survey were skewed towards full health, as reported in chapter 3 (section 3.8 and table 3.3). More data for subjects with poorer HRQL would have strengthened the robustness of the scoring algorithm across the full spectrum of health and disease, but selecting a greater proportion of patients with poorer HRQL would have required interviews with patients in hospital or in the terminal phase of their illness. Hospitalised patients may be too unwell for cognitively demanding interviews. Patients in the terminal phase of their illness may find that questions about life and death are too confronting. Another approach would have directed patients to assign utilities to hypothetical health states across the full range, rather than their current health state. The disadvantage of eliciting utilities for hypothetical health states is that the valuations may differ from those for experienced health states [59].

The utilities assigned in the valuation survey were highly skewed towards full health, and half of the subjects assigned a utility of one (section 3.9 in chapter 3). This implies that many were unwilling to trade any survival time for an improvement in HRQL. One explanation for this finding is that the subjects truly valued survival more highly than HRQL. This is supported by 58% of subjects reporting their health as being 'excellent' or 'good', despite being recruited to the study because they had advanced cancer and impaired HRQL (table 3.2 of chapter 3). Another explanation is that many failed to comprehend the task, or objected to it. Assigning utilities by time trade-off interview may be too demanding for some patients with advanced cancer. Patients may have as much difficulty remembering full health as lay people have imagining severely impaired health [59]. Another is that they may adjust their valuations because of adaptation and response shift (section 2.4.3 of chapter 2). However the strength of using experienced patients rather than lay people to assign utilities is that they have a proper understanding of what it is like to live in that health state.

A number of direct utility-based scaling methods could have been used to assign utilities in the valuation survey. The strengths and limitations of each were described in chapter 2 (section 2.4.3). In summary, the time trade-off method was used because it is practical, reliable and has empirical validity. One alternative is the standard gamble, which was used to derive scoring algorithms for the utility indices of the Health Utilities Index and SF-6D (table 2.2 of chapter 2). The standard gamble is the criterion method for direct utility-based scaling because it elicits utilities under conditions of uncertainty, but the time trade-off is generally accepted as having similar empirical properties to the standard gamble, and may be easier to complete. Martin et al. developed methods to convert time trade-off utilities to standard gamble utilities where they differed. In this study, this would further increase the utility scores [49]. This conversion was not used in this study because utilities were already strongly skewed towards full health. Another alternative to the time trade-off is the rating scale. This is an inferior method for eliciting utilities on both theoretical and empirical grounds. Data presented in chapter 3 confirmed that values elicited from patients about their current health status were significantly lower when elicited with a rating scale than with a time trade-off interview (table 3.3 versus section 3.9).

The data in the valuation survey was used to produce the scoring algorithm for the utility index. The next section describes the merits of the approach taken.

8.4.3 Producing the scoring algorithm

A statistical inference approach was used to produce the scoring algorithm for the utility index. The alternatives presented in chapter 2 (section 2.6.3) were the holistic approach, and the multi-attribute utility theory approach. The strength of the statistical inference approach over these alternatives derives from the type of health states that need to be valued. The holistic approach requires valuation of every conceivable health state, and the multi-attribute utility theory approach requires valuation of 'corner health states' with impairment of only one attribute. It would be difficult to recruit subjects with experience across the full spectrum of health states required. The statistical inference approach required only a representative sample of health states to be valued.

Two standard statistical inference approaches were described in chapter 2 (sections 2.5.1 and 2.6.3). The global health preference approach uses a single-item global scale, and the multi-attribute health preference approach uses multiple items. The strength of a single-item global scale is that it provides a single unified assessment of global HRQL; the strength of multiple items is that they improve responsiveness and the descriptive richness of the index, as discussed in chapter 2 (section 2.5.2). The scoring algorithm for the utility index used a novel statistical approach that combined a single-item global scale with multi-item subscales, incorporating the strengths of each approach. This enabled the index to be optimised to better reflect the attitudes of patients in specific clinical contexts, as discussed in the next section.

There are a number of limitations of our approach. First, it is more complex to understand and implement. Second, the weight that is applied to the multi-item scales (subset index) was low and remarkably constant (about 0.33), such that they contribute little to the utility index. Third, the greater precision of the HRQL index compared to the health status thermometer could potentially reflect ignorance of uncertainty in its development. Fourth, the scores derived by the utility index in different contexts may be less comparable if the optimal algorithms differ across different contexts. All these issues warrant future research.

The utility index scores derived by the resultant scoring algorithm were sufficiently similar to directly assigned, time trade-off utilities for group comparisons, as reported in chapter 5, but the algorithm did not accurately predict utilities for individuals. An alternative approach could have incorporated other factors that influence utilities for individuals into the scoring algorithm, such as attitudes to risk, social and demographic factors that were outlined in chapter 2 (section 2.4). However, incorporating these factors is unlikely to yield precise estimates for individuals, because of the inherent variability of utilities. Another alternative approach would have compared the power transformation to a plateau model that took into account the large number of respondents with a utility of 1, as discussed in chapter 5 (section 5.6).

8.4.4 Optimising the scoring algorithm in specific clinical contexts

The scoring algorithm was optimised for trials of chemotherapy in early and advanced breast cancer. The limitation of the approach taken to optimising the weights is that it is data-driven, with weights assigned to each subscale based on their correlations with the health status thermometer. An alternative approach would be to optimise the algorithm by adjusting subscale weights based on their correlations with direct time trade-off utilities. This may better reflect the attitudes of respondents but requires substantial additional work. Another alternative or complementary approach would establish that the weights truly reflect the relative importance that patients attach to aspects of HRQL by qualitative interviews or questionnaires, again necessitating substantial additional effort.

8.4.5 Validation using related measures of HRQL

The utility index was validated by application to data from clinical trials of cancer treatments. The strength of this approach is that it used the type of data that the index was designed to analyse. It was validated as a measure of HRQL by comparison with related measures, with strong preliminary evidence of validity, sensitivity and responsiveness. An alternative or complementary approach would have focussed on establishing the validity of the utility index by comparison with directly elicited utilities in additional valuation surveys, or by dividing a larger valuation survey into a ‘development’ dataset where the algorithm was produced, and a ‘validation’ dataset where its predictive ability was tested. The responsiveness and sensitivity of the

utility index compared with those from other utility-based instruments could have also been tested.

8.4.6 Application to treatment comparisons in randomised trials

The utility index was applied to treatment comparisons of chemotherapy regimens for early and advanced breast cancer. These treatment comparisons were ideal because they are clinically relevant, and the utility scores that were generated are being used to inform decisions about these treatments in terms of QALYs in ongoing analyses. One limitation is that little difference in utility was found between treatment groups, as reported in section 8.3. This probably reflects the genuine similarities in overall HRQL between treatment groups, but could reflect the limited sensitivity of the utility index. A complementary approach would have applied the index in other types of cancer, and other treatment modalities such as surgery or supportive care, and compared it to scores from other utility indices, to better understand its applicability and sensitivity.

8.5 Practical implications of research for future trials

The general aim of this thesis (section 8.1) proposed the ideal characteristics of a utility index for use in clinical trials to inform clinical decision-making about cancer treatments. Such a utility index would reflect the perspective of patients with cancer, be optimally weighted for comparisons within a specific disease and treatment context, and be feasible for use in clinical trials. The previous sections of this chapter reported how the work reported in this thesis successfully achieved these ideal characteristics. This section considers the practical implications of such a utility index for the conduct and evaluation of future clinical trials.

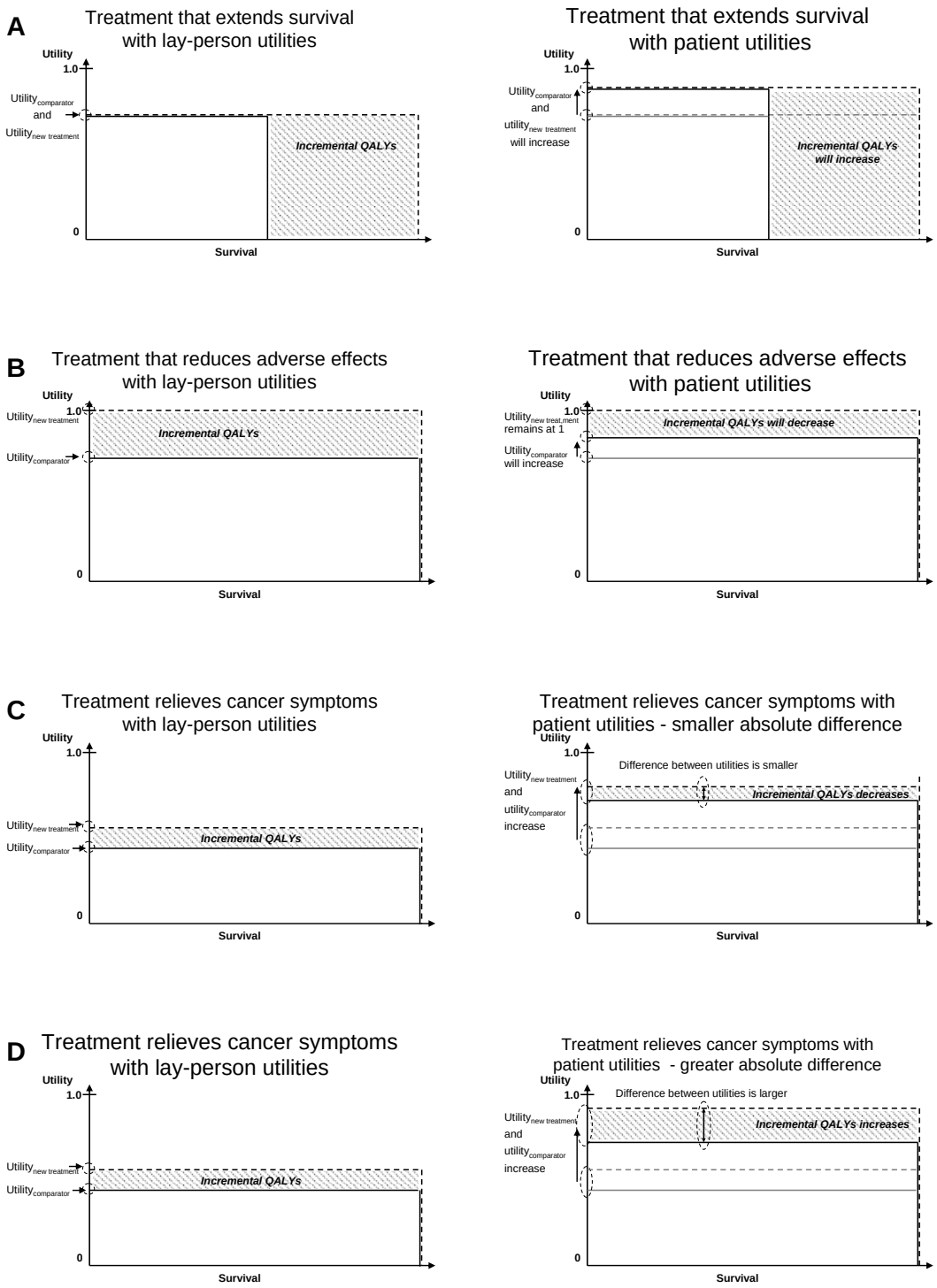
8.5.1 Reflecting the perspective of patients with cancer

The utility index was intended to reflect the perspective of patients, whereas indices for most utility-based instruments are designed to reflect the perspective of lay people. Because patient-based utilities are typically higher than community-based utilities, their use is likely to redirect priorities away from treatments that improve HRQL and towards treatments that extend life [59]. The effects are described in table 8.2, and are explained in the following paragraphs.

Table 8.2 Effects of patient-based utilities on incremental benefit of interventions

Treatment effect	**Examples of intervention and comparator**	**Effect of patient-based utilities on incremental benefit**
Extend survival	1. Curative surgery versus nil 2. Life-extending chemotherapy versus nil	More favourable
Reduce adverse effects of treatment	1. Surgery that avoids colostomy versus surgery with colostomy 2. Mastectomy with breast reconstruction versus mastectomy 3. Adjuvant hormonal therapy without side effects versus toxic adjuvant hormonal therapy	Less favourable
Relieve cancer symptoms	Palliative chemotherapy versus nil Palliative radiotherapy versus nil	Variable

Figure 8.1 Effects of patient-based utilities on incremental benefit of interventions



Utilities valued by patients, rather than by lay people, are likely to make the incremental benefit of an intervention over a comparator that extends survival but does not alter HRQL seem more favourable. In this context, patient-based utilities assign higher utilities and more QALYs to both the intervention and the comparator. Because the utility of the intervention and the comparator is the same but the survival differs, the effect is greater incremental QALYs. The effect is illustrated in figure 8.1 (A). This explains why cancer treatments that extend survival, such as life-prolonging chemotherapy or curative surgery, seem more effective and cost-effective when patient-based utilities are used.

In contrast, utilities valued by patients, rather than by lay people are likely to make the incremental benefit of an intervention over a comparator that restores perfect health but does not alter survival seem less favourable. In this context, patient-based utilities assign a higher utility and more QALYs to the comparator, but the utility of the intervention is fixed at one and its QALY is unchanged. Because the survival of the intervention and the comparator is the same, the effect is smaller incremental gains in QALYs between the intervention and the comparator [59, 71]. The effect is illustrated in figure 8.1 (B). This explains why treatments that reduce adverse effects of treatments but do not alter survival, such as breast reconstruction following mastectomy, or avoidance of colostomy, seem less effective and less cost-effective when patient-based utilities are used.

Utilities valued by patients, rather than by lay people, can have variable consequences on the effectiveness and cost-effectiveness on interventions that incrementally improve HRQL, but do not restore HRQL to full health or extend survival. Patient-based utilities assign higher utilities to both the intervention and the comparator. Because survival of the intervention and the comparator is the same, the gain in QALYs is determined only by the absolute difference in utilities between them. If the absolute difference in utilities is smaller, then the effect is less gains in QALYs and diminished effectiveness and cost-effectiveness, as illustrated in figure 8.2 (C) [209-210]. This tends to occur for less severe health states, because of a phenomenon called valuation compression where the valuations for less severe health states are all compressed near the top of the scale [59]. Conversely, if the absolute difference is greater then the effect is more gains in QALY and enhanced

effectiveness and cost-effectiveness, as illustrated in figure 8.2 (D). This tends to occur for more severe health states, because patients are better than lay people at distinguishing between them [59]. If the absolute difference in utilities is the same then the effectiveness and cost-effectiveness are unchanged.

In summary, the use of utilities valued by patients compared to lay people in analyses of quality-adjusted survival and cost-effectiveness will favour treatments that prolong survival over treatments that improve HRQL. This finding requires validation by correlation with actual decision-making by patients. The implication is that many patients with cancer are unwilling to trade reductions in survival times for improvements in HRQL. It should be noted that utilities valued by patients rather than lay people may not always have a significant impact on decisions in specific contexts. Chapman et al reported that varying utilities substantially does not alter decisions in a sizeable proportion of cost-utility analyses [211].

The next section considers the implications of another characteristic of the utility index: its optimal weighting.

8.5.2 Optimal weighting

The utility index was designed to be optimally weighted for comparisons within a specific disease and treatment context. This section considers the implications of using an optimally-weighted index in clinical trials in terms of the comparability of scores across trials, precision of estimates, information about trade-offs, and expertise required for analyses.

The optimally-weighted scores on the utility index should better reflect the perspective of the respondents, because it gives additional weight to the aspects of HRQL that are more strongly correlated with global HRQL. The main disadvantage is that the utility scores may be less comparable in different contexts, because the weights may differ. This is not a limitation for evaluations of treatments within a specific clinical trial to inform clinical decision-making, because the comparisons are restricted to a single clinical context. However it does limit application of the method to studies that include diverse populations, or for comparing utilities and QALYs from one study to another.

Optimal weighting could give more precise estimates of clinically important differences in utility between patient groups, because the index is focussed on those aspects of HRQL that are most relevant to those patients. A more precise utility index has implications for decision-making about cancer treatments. It will reduce the uncertainty around the incremental effectiveness and cost-effectiveness of treatments in sensitivity analyses, because it is more responsive to small but meaningful effects of cancer treatments [172]. It will reduce the sample size required to detect a given difference with a given level of precision [172]. A more precise utility index may also make an intervention versus a comparator seem more favourable if it detects greater differences in utilities between treatment groups. As discussed in the previous section and figure 8.1 (D), a larger absolute difference in utilities and QALYs between an intervention and its comparator will enhance its incremental effectiveness and cost-effectiveness, but a smaller absolute difference will diminish it.

Optimal weighting may help researchers and clinicians better understand the relative importance of different aspects of HRQL in a specific clinical context. The implications are illustrated by table 8.1. For example, distresses were assigned the greatest weight and self-care the least weight in our studies. The implication is that clinicians and researchers should pay greater attention to relieving distresses due to physical and psychological symptoms in these contexts, rather than focussing on self-care, because self-care wasn't as severely impaired in the studies. The effects of optimising the weights to reflect the preferences of individual patients on trial results can also be tested in sensitivity analyses.

The approach to optimal weighting requires familiarity with the methods developed in this thesis. The work reported in this thesis has shown the feasibility and advantages of optimal weighting. However, implementation of this approach requires additional work for statisticians.

8.5.3 Feasibility of use in clinical trials

The feasibility of the utility index for use in clinical trials of cancer treatments was supported by the high completion rate of the UBQ-C in the clinical trials evaluated in this thesis, as reported in chapter 3, and the ability to derive utility scores from the UBQ-C in each trial, as reported in chapters 6 and 7.

A utility index that is feasible for use in clinical trials should increase the pool of utility scores available for treatment comparisons. A current limitation of evaluating clinical trials in terms of utilities and QALYs is the limited number of specific health states for which utility data is available. The utility index developed in this thesis should promote the use of these analyses to evaluate treatments in randomised trials. The analyses can be used to inform clinical decision-making, and provide an alternate perspective to lay person-based utilities to inform evaluation of the incremental cost-effectiveness of treatments to inform funding and policy decisions.

Deriving utility scores with the utility index developed in this thesis is more feasible than eliciting utilities directly with a standard gamble or time trade-off interview. The direct approach is not practical for cancer patients participating in clinical trials because the procedure is resource-intensive for researchers, difficult for patients, and potentially distressing for those with severe symptoms or side effects, or at the end of life [59], as described in chapter 2 (section 2.4.4). The utility index developed in this thesis is also more efficient in clinical trials of cancer treatments than standard generic utility-based instruments such as the EuroQol EQ-5D or Health Utilities Index (HUI3), because demands on patients and trial staff are reduced by using a single instrument that gives both a cancer-specific, profile-based, description of HRQL, and a utility score.

8.6 Priorities for ongoing and future research

Current work is further establishing the validity of this utility index as a HRQL measure by testing its sensitivity and responsiveness in other contexts such as chemotherapy for advanced colorectal cancer [150] and chemoprevention for breast cancer [126]. Current work is also comparing its measurement properties with those of the EuroQol EQ-5D, which is one of the most commonly used generic utility-based instruments [150]. Publications relating to this work are in progress.

Future work could further establish the validity of the utility index by comparing its scores with utilities directly elicited by time trade-off interview about patients' current health in new samples. Sensitivity analyses should test if optimisation of the scoring algorithm in different clinical contexts makes a meaningful difference to the utility scores, and to their sensitivity and responsiveness.

Application of the utility index to inform decision-making will be an important extension of this work. Comparison of treatments in terms of incremental QALYs and incremental cost per QALYs is planned in four clinical contexts where data about survival, HRQL and costs have been collected: the early and advanced cancer trials presented in this thesis, chemotherapy for advanced colorectal cancer [150], and chemoprevention for breast cancer [126]. Publications relating to analyses of these randomised controlled trials in terms of utilities and QALYs are planned or in progress.

Planned sensitivity analyses comparing patient-based utilities derived from the UBQ-C with lay person-based utilities derived from the EQ-5D will determine whether differences in utilities make a meaningful difference to the outcomes of decisions in each context. Sensitivity analyses will also explore the effect on treatment recommendations of optimising the scoring algorithm in different clinical contexts.

Future work should look at adaptations of the methods used to derive the utility index. One adaptation would be to derive the utility index from a weighted combination of items rather than subscales. This would provide additional information about the trade-offs between different aspects of HRQL (section 8.4.1).

Another adaptation would be to derive a utility index for the UBQ-C based on the perspective of lay people, by repeating the valuation survey in a general population. This would facilitate the comparison of patient-based and lay-people based utilities with a single instrument. Finally, the methodological approach could be applied to derive a utility index for other HRQL instruments that are relevant to cancer or other diseases.

8.7 Contribution to knowledge and practical significance

The work presented in this thesis has developed a scoring algorithm that converts the responses from a simple, self-rated cancer-specific questionnaire about health-related quality of life to a utility index. The index can be used to generate utility scores and quality-adjusted life-years in clinical trials. It enables the evaluation of the net effect of treatments on health-related quality of life (accounting for trade-offs between disparate aspects); the evaluation of the net benefit of treatments (accounting for trade-offs between quality of life and survival); and an alternate perspective for comparing the incremental cost-effectiveness of treatments (accounting for trade-offs between net benefit and costs).

The practical significance of this work is to facilitate the integration of data about health-related quality of life with traditional trial endpoints such as survival and tumour response. This will better inform interpretation of clinical trials that are used to inform clinical decision-making, and provide an alternate viewpoint for economic decision-making. Broadly, it will help patients, clinicians and health funders make better decisions about cancer treatments by considering effects on both length and quality of life.

9. Bibliography

[1] Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. Ann Intern Med. 1993 Apr 15;118(8):622-9.

[2] Fayers P, Machin D. Quality of Life: The Assessment, Analysis and Interpretation of Patient-reported Outcomes. 2nd ed. Chichester, England: John Wiley & Sons; 2007.

[3] Spilker B. Quality of life and pharmacoeconomics in clinical trials. 2nd ed. ed. Philadelphia: Raven Press; 1996.

[4] Naughton MJ, Shumaker SA, Naughton MJ, Shumaker SA. The case for domains of function in quality of life assessment. Quality of Life Research. 2003;12 Suppl 1:73-80.

[5] McDowell I. Measuring health : a guide to rating scales and questionnaires. 3rd ed. New York: Oxford University Press; 2006.

[6] Schumacher M, Olschewski M, Schulgen G. Assessment of quality of life in clinical trials. Stat Med. 1991 Dec;10(12):1915-30.

[7] Leplege A, Hunt S. The problem of quality of life in medicine. Jama. 1997 Jul 2;278(1):47-50.

[8] Fayers PM. Quality-of-life measurement in clinical trials--the impact of causal variables. J Biopharm Stat. 2004 Feb;14(1):155-76.

[9] Drummond MF. Methods for the economic evaluation of health care programmes. 3rd ed. Oxford: Oxford University Press; 2005.

[10] Lipscomb J, Reeve BB, Clauser SB, Abrams JS, Bruner DW, Burke LB, et al. Patient-reported outcomes assessment in cancer trials: taking stock, moving forward. J Clin Oncol. 2007 Nov 10;25(32):5133-40.

[11] Fayers PM, Hays RD. Assessing quality of life in clinical trials : methods and practice. 2nd ed. ed. Oxford: Oxford University Press; 2005.

[12] Martin AJ, Greatorex V, Simes RJ. Preliminary results on the reliability and validity of the UBQ-C(ancer) items: NHMRC Clinical Trials Centre, University of Sydney1996.

[13] Martin AJ, Greatorex V, Simes RJ. Towards a utility-based assessment for cancer patients: Reliability and validity of the UBQ-C(ancer) items (Abstract P81). Controlled Clinical Trials. 1997;18(3, Supplement 1):S160-S.

[14] Sloan JA, Loprinzi CL, Kuross SA, Miser AW, O'Fallon JR, Mahoney MR, et al. Randomized comparison of four tools measuring overall quality of life in patients with advanced cancer. J Clin Oncol. 1998 Nov;16(11):3662-73.

[15] Spitzer WO, Dobson AJ, Hall J, Chesterman E, Levi J, Shepherd R, et al. Measuring the quality of life of cancer patients : A concise QL-Index for use by physicians. Journal of Chronic Diseases. 1981;34(12):585-97.

[16] Green C, Brazier J, Deverill M. Valuing health-related quality of life. A review of health state valuation techniques. Pharmacoeconomics. 2000 Feb;17(2):151-65.

[17] Revicki DA, Kaplan RM. Relationship between psychometric and utility-based approaches to the measurement of health-related quality of life. Quality of Life Research. 1993;2(6):477-87.

[18] Sloan JA, Aaronson N, Cappelleri JC, Fairclough DL, Varricchio C. Assessing the clinical significance of single items relative to summated scores. Mayo Clinic proceedings. 2002 May;77(5):479-87.

[19] Youngblut JM, Casper GR. Single-item indicators in nursing research. Res Nurs Health. 1993 Dec;16(6):459-65.

[20] de Boer AGEM, van Lanschot JJB, Stalmeier PFM, van Sandick JW, Hulscher JBF, de Haes JCJM, et al. Is a single-item visual analogue scale as valid, reliable and responsive as multi-item scales in measuring quality of life? Quality of Life Research. 2004;13(2):311-20.

[21] Somerfield MR, Sloan J, Loprinzi C. Wherefore Global Quality-of-Life Assessment? J Clin Oncol. 1999 Feb 1;17(2):730.

[22] Bowling A. Just one question: If one question works, why ask several? J Epidemiol Community Health. 2005 May 1;59(5):342-5.

[23] Ware JEJ, Gandek B. Overview of the SF-36 Health Survey and the International Quality of Life Assessment (IQOLA) Project. Journal of Clinical Epidemiology. 1998;51(11):903-12.

[24] Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, et al. The European 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 March 3;85(5):365-76.

[25] Fayers P, Aaronson N, Bjordal K, et al. EORTC QLQ-C30 Scoring Manual. 3rd ed. Brussels, Belgium: EORTC Publications; 2001.

[26] Cella DF, Tulsky DS, Gray G, Sarafian B, Linn E, Bonomi A, et al. The Functional Assessment of Cancer Therapy scale: development and validation of the general measure. J Clin Oncol. 1993 Mar;11(3):570-9.

[27] Olschewski M, Schumacher M. Statistical analysis of quality of life data in cancer clinical trials. Stat Med. 1990 Jul;9(7):749-63.

[28] Streiner DL, Norman GR. Health measurement scales : a practical guide to their development and use. 3rd ed. Oxford ; New York: Oxford University Press; 2003.

[29] Cox DR, Fitzpatrick R, Fletcher AE, Gore SM, Spiegelhalter DJ, Jones DR. Quality-of-Life Assessment: Can We Keep It Simple? Journal of the Royal Statistical Society Series A (Statistics in Society). 1992;Vol. 155(3):353-93.

[30] Wainer H. Estimating coefficients in linear models: It don't make no nevermind. Psychol Bull. 1976;83(2):213-7.

[31] Buchanan DR, O'Mara AM, Kelaghan JW, Sgambati M, McCaskill-Stevens W, Minasian L. Challenges and recommendations for advancing the state-of-the-science of quality of life assessment in symptom management trials. Cancer. 2007;110(7):1621-8.

[32] Lumley T, Simes RJ, Gebski V, Hudson HM. Combining components of quality of life to increase precision and evaluate trade-offs. Stat Med. 2001 Nov 15;20(21):3231-49.

[33] Johnson FR, Hauber AB, David O, Ming-Ann H, John C, Catherine C-M. Are Chemotherapy Patients' HRQoL Importance Weights Consistent with Linear Scoring Rules? A Stated-choice Approach. Quality of Life Research. 2006;V15(2):285-98.

[34] Schipper H, Clinch J, McMurray A, Levitt M. Measuring the quality of life of cancer patients: the Functional Living Index-Cancer: development and validation. J Clin Oncol. 1984 May 1;2(5):472-83.

[35] Bagust A, Barraza-Llorens M, Philips Z. Deriving a compound quality of life measure from the EORTC-QLQ-C30/LC13 instrument for use in economic evaluations of lung cancer clinical trials. European Journal of Cancer. 2001;37(9):1081-8.

[36] Berger ML, Bingefors K, Hedblom EC, Pashos CL, Torrance GW, Smith MD. Health Care Cost, Quality, and Outcomes: ISPOR Book of Terms. Lawrenceville, NJ: International Society for Pharmacoeconomics and Outcomes Research; 2003.

[37] Brazier J, Green C, McCabe C, Stevens K. Use of visual analog scales in economic evaluation. Expert Rev Pharmacoeconomics Outcomes Res. 2003;3(3):293–302.

[38] Martin AJ, Lumley TS, Simes RJ. Incorporating trade-offs in quality of life assessment (Book Chapter). In: Spilker B, editor. Quality of life and pharmacoeconomics in clinical trials 2nd ed. Philadelphia: Lippincott-Raven; 1996. p. 403-12.

[39] Froberg DG, Kane RL. Methodology for measuring health-state preferences--I: Measurement strategies. J Clin Epidemiol. 1989;42(4):345-54.

[40] Torrance GW, Feeny D. Utilities and quality-adjusted life years. Int J Technol Assess Health Care. 1989;5(4):559-75.

[41] Richardson J. Cost utility analysis: what should be measured? Soc Sci Med. 1994 Jul;39(1):7-21.

[42] Froberg DG, Kane RL. Methodology for measuring health-state preferences--II: Scaling methods. J Clin Epidemiol. 1989;42(5):459-71.

[43] Torrance GW. Measurement of health state utilities for economic appraisal. J Health Econ. 1986 Mar;5(1):1-30.

[44] Torrance GW, Thomas WH, Sackett DL. A utility maximization model for evaluation of health care programs. Health Serv Res. 1972 Summer;7(2):118-33.

[45] Feeny D. Preference-based measures: utility and quality-adjusted life years (Book Chapter). In: Fayers PM, Hays RD, editors. Assessing quality of life in clinical trials : methods and practice 2nd ed. ed. Oxford: Oxford University Press; 2005. p. 405-29.

[46] Kaplan RM. Utility assessment for estimating quality-adjusted life years. In: Sloan FA, editor. Valuing health care : costs, benefits, and effectiveness of

pharmaceuticals and other medical technologies. Cambridge ; New York: Cambridge University Press; 1995. p. pp 31-60.

[47] Bravata DM, Nelson LM, Garber AM, Goldstein MK. Invariance and inconsistency in utility ratings. Med Decis Making. 2005 Mar-Apr;25(2):158-67.

[48] Nord E. Methods for quality adjustment of life years. Social Science & Medicine. [Review]. 1992 Mar;34(5):559-69.

[49] Martin AJ, Glasziou PP, Simes RJ, Lumley T. A comparison of standard gamble, time trade-off, and adjusted time trade-off scores. Int J Technol Assess Health Care. 2000 Winter;16(1):137-47.

[50] Dolan P, Gudex C, Kind P, Williams A. Valuing health states: a comparison of methods. J Health Econ. 1996 Apr;15(2):209-31.

[51] Torrance G. Social preferences for health states: an empirical evaluation of three measurement techniques. Socio-Economic Planning Sciences. 1976;10:129-36.

[52] Revicki DA. Relationship between health utility and psychometric health status measures. Med Care. 1992 May;30(5 Suppl):MS274-82.

[53] Dolan P, Kahneman D. Interpretations Of Utility And Their Implications For The Valuation Of Health. The Economic Journal. 2008;118(525):215-34.

[54] Simes RJ. Application of statistical decision theory to treatment choices: implications for the design and analysis of clinical trials. Stat Med. 1986 Sep-Oct;5(5):411-20.

[55] Froberg DG, Kane RL. Methodology for measuring health-state preferences--III: Population and context effects. J Clin Epidemiol. 1989;42(6):585-92.

[56] Ratcliffe J, Brazier J, Palfreyman S, Michaels J. A comparison of patient and population values for health states in varicose veins patients. Health Economics. 2007;16(4):395-405.

[57] Richardson J, Nord E. The importance of Perspective in the Measurement of Quality-adjusted Life Years. Med Decis Making. 1997 Feb 1;17(1):33-41.

[58] Ubel PA, Loewenstein G, Jepson C. Whose quality of life? A commentary exploring discrepancies between health state evaluations of patients and the general public. Qual Life Res. 2003 Sep;12(6):599-607.

[59] Brazier J, Akehurst R, Brennan A, Dolan P, Claxton K, McCabe C, et al. Should Patients Have a Greater Role in Valuing Health States? Applied Health Economics and Health Policy. 2005;4(4):201-8.

[60] Elkin EB, Cowen ME, Cahill D, Steffel M, Kattan MW. Preference Assessment Method Affects Decision-Analytic Recommendations: A Prostate Cancer Treatment Example. Med Decis Making. 2004 Oct 1;24(5):504-10.

[61] Stiggelbout AM, de Haes JCJM. Patient Preference for Cancer Therapy: An Overview of Measurement Approaches. J Clin Oncol. 2001 January 1, 2001;19(1):220-30.

[62] Stiggelbout AM, de Vogel-Voogt E. Health State Utilities: A Framework for Studying the Gap Between the Imagined and the Real. Value Health. 2008;11(1):76-87.

[63] Osoba D, Hsu M-A, Copley-Merriman C, Coombs J, Johnson FR, Hauber B, et al. Stated Preferences of Patients with Cancer for Health-related Quality-of-life (HRQOL) Domains During Treatment. Quality of Life Research. 2006;15(2):273-83.

[64] Rose M, Scholler G, Klapp BP, Bernheirn JL. Weighting dimensions in “generic” QOL questionnaires by Anamnestic Comparative Self-assessment: different weights in different diseases (Abstract). Quality of Life Research. [Abstract]. 1998;7:655.

[65] Wittenberg E, Halpern E, Divi N, Prosser LA, Araki SS, Weeks JC. The effect of age, race and gender on preference scores for hypothetical health states. Qual Life Res. 2006 May;15(4):645-53.

[66] Fryback DG. A US Valuation of the EQ-5D. Medical Care. 2005 March;43(3):199.

[67] Johnson JA, Luo N, Shaw JW, Kind P, Coons SJ. Valuations of EQ-5D Health States: Are the United States and United Kingdom Different? Medical Care. 2005;43(3):221.

[68] Dowie J. Decision validity should determine whether a generic or condition-specific HRQOL measure is used in health care decisions. Health Economics. 2002;11(1):1-8.

[69] Australian Government Department of Health and Ageing. Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee (Version 4.0). Canberra2006.

[70] National Institute for Clinical Excellence (NICE). Guide to the Methods of Technology Appraisal. London: NICE; 2004.

[71] Russell LB, Gold MR, Siegel JE, Daniels N, Weinstein MC. The role of cost-effectiveness analysis in health and medicine. Panel on Cost-Effectiveness in Health and Medicine. JAMA. 1996 Oct 9;276(14):1172-7.

[72] Froberg DG, Kane RL. Methodology for measuring health-state preferences--IV: Progress and a research agenda. J Clin Epidemiol. 1989;42(7):675-85.

[73] Gold MR. Cost-effectiveness in health and medicine. New York: Oxford University Press; 1996.

[74] Sullivan PW, Lawrence WF, Ghushchyan V, Sullivan PW, Lawrence WF, Ghushchyan V. A national catalog of preference-based scores for chronic conditions in the United States. Medical Care. 2005 Jul;43(7):736-49.

[75] Glasziou PP, Cole BF, Gelber RD, Hilden J, Simes RJ. Quality adjusted survival analysis with repeated quality of life measures. Stat Med. 1998 Jun 15;17(11):1215-29.

[76] Au H-J, Golmohammadi K, Younis T, Verma S, Chia S, Fassbender K, et al. Cost-effectiveness analysis of adjuvant docetaxel, doxorubicin, and cyclophosphamide (TAC) for node-positive breast cancer: modeling the downstream effects. Breast Cancer Research and Treatment (Online First). 2008;doi 10.1007/s10549-008-0034-1.

[77] Takeda AL, Jones J, Loveman E, Tan SC, Clegg AJ. The clinical effectiveness and cost-effectiveness of gemcitabine for metastatic breast cancer: a systematic review and economic evaluation. Health Technol Assess. 2007 May;11(19):1-80.

[78] Locker GY, Mansel R, Cella D, Dobrez D, Sorensen S, Gandhi SK. Cost-effectiveness analysis of anastrozole versus tamoxifen as primary adjuvant therapy for postmenopausal women with early breast cancer: a US healthcare system perspective. The 5-year completed treatment analysis of the ATAC ('Arimidex', Tamoxifen Alone or in Combination) trial. Breast Cancer Res Treat. 2007 Jan 24;106(2):229-38.

[79] Kurian AW, Thompson RN, Gaw AF, Arai S, Ortiz R, Garber AM. A Cost-Effectiveness Analysis of Adjuvant Trastuzumab Regimens in Early HER2/neu-Positive Breast Cancer. J Clin Oncol. 2007 Feb 20;25(6):634-41.

[80] van den Brink M, van den Hout WB, Stiggelbout AM, Klein Kranenbarg E, Marijnen CAM, van de Velde CJH, et al. Cost-Utility Analysis of Preoperative Radiotherapy in Patients With Rectal Cancer Undergoing Total Mesorectal Excision: A Study of the Dutch Colorectal Cancer Group. J Clin Oncol. 2004 Jan 15;22(2):244-53.

[81] Eckermann SD, Martin AJ, Stockler MR, Simes RJ. The benefits and costs of tamoxifen for breast cancer prevention. Australian and New Zealand journal of public health. 2003;27(1):34-40.

[82] Pickard AS, De Leon MC, Kohlmann T, Cella D, Rosenbloom S. Psychometric comparison of the standard EQ-5D to a 5 level version in cancer patients. Med Care. 2007 Mar;45(3):259-63.

[83] Remak E, Charbonneau C, Negrier S, Kim ST, Motzer RJ. Economic evaluation of sunitinib malate for the first-line treatment of metastatic renal cell carcinoma. J Clin Oncol. 2008 Aug 20;26(24):3995-4000.

[84] Mallick R, Hudes G, Levy DE. Clinically important differences in quality-adjusted survival in patients with advanced renal cell carcinoma receiving first-line treatment with temsirolimus or interferon-alpha (Abstract PCN64). Value in Health, Proceedings of International Society for Pharmacoeconomic and Outcomes Research Tenth Annual European Congress, Dublin. 2007;10(6):A342.

[85] Bernhard J, Zahrieh D, Zhang JJ, Martinelli G, Basser R, Hurny C, et al. Quality of life and quality-adjusted survival (Q-TWiST) in patients receiving dose-intensive or standard dose chemotherapy for high-risk primary breast cancer. Br J Cancer. 2008 Nov 27;98(1):25-33.

[86] Bernhard J, Zahrieh D, Coates AS, Gelber RD, Castiglione-Gertsch M, Murray E, et al. Quantifying trade-offs: quality of life and quality-adjusted survival in a randomised trial of chemotherapy in postmenopausal patients with lymph node-negative breast cancer. Br J Cancer. 2004 Nov 29;91(11):1893-901.

[87] Grann VR, Jacobson JS, Thomason D, Hershman D, Heitjan DF, Neugut AI. Effect of Prevention Strategies on Survival and Quality-Adjusted Survival of Women With BRCA1/2 Mutations: An Updated Decision Analysis. J Clin Oncol. 2002 May 15;20(10):2520-9.

[88] Sommers BD, Beard CJ, D'Amico AV, Dahl D, Kaplan I, Richie JP, et al. Decision analysis using individual patient preferences to determine optimal treatment for localized prostate cancer. Cancer. 2007;110(10):2210 - 7.

[89] Kilbridge KL, Cole BF, Kirkwood JM, Haluska FG, Atkins MA, Ruckdeschel JC, et al. Quality-of-life-adjusted survival analysis of high-dose

adjuvant interferon alpha-2b for high-risk melanoma patients using intergroup clinical trial data. J Clin Oncol. 2002 Mar 1;20(5):1311-8.

[90] Kopec JA, Willison KD. A comparative review of four preference-weighted measures of health-related quality of life. Journal of Clinical Epidemiology. 2003;56(4):317-25.

[91] Shaw J, Pickard A, Lin H, Cella D, Trask P. Evaluation of a theory of global health preference formation. Value in Health, ISPOR Thirteenth Annual International Meeting Proceedings. 2008;11(3):A14.

[92] Torrance GW, Feeny D, Furlong W. Visual Analog Scales: Do They Have a Role in the Measurement of Preferences for Health States? Med Decis Making. 2001 July 1;21(4):329-34.

[93] Torrance GW, Feeny DH, Furlong WJ, Barr RD, Zhang Y, Wang Q. Multiattribute utility function for a comprehensive health status classification system. Health Utilities Index Mark 2. Med Care. 1996 Jul;34(7):702-22.

[94] Hurny C, van Wegberg B, Bacchi M, Bernhard J, Thurlimann B, Real O, et al. Subjective health estimations (SHE) in patients with advanced breast cancer: an adapted utility concept for clinical trials. Br J Cancer. 1998 Mar;77(6):985-91.

[95] Pickard AS, Shaw JW, Lin H-W, Trask PC, Aaronson N, Lee TA, et al. A Patient-based Utility Measure of Health for Clinical Trials of Cancer Therapy (Abstract 6529). Journal of Clinical Oncology, ASCO Annual Meeting Proceedings. 2008;25(15S):344S.

[96] Coens C, Bottomley A, Efficace F, Flechtner H, Aaronson N. Variability and Sample Size Requirements for Health-Related Quality-of-Life Measures: Understanding the Challenges Facing Investigators. J Clin Oncol. 2005 Nov 20;23(33):8541-2.

[97] Brazier J, Czoski-Murray C, Roberts J, Brown M, Symonds T, Kelleher C. Estimation of a Preference-Based Index from a Condition-Specific Measure: The King's Health Questionnaire. Med Decis Making. 2008 Feb 1;28(1):113-26.

[98] Dobrez D, Cella D, Pickard AS, Lai J-S, Nickolov A. Estimation of Patient Preference-Based Utility Weights from the Functional Assessment of Cancer Therapy-General. Value Health. 2007;10(4):266-72.

[99] Noyes K, Dick AW, Holloway RG. The Implications of Using US-Specific EQ-5D Preference Weights for Cost-Effectiveness Evaluation. Medical Decision Making. 2007;27(3):327-34.

[100] Shaw JW, Johnson JA, Coons SJ. US valuation of the EQ-5D health states: development and testing of the D1 valuation model. Med Care. 2005 Mar;43(3):203-20.

[101] Glasziou P, Alexander J, Beller E, Clarke P. Which health-related quality of life score? A comparison of alternative utility measures in patients with Type 2 diabetes in the ADVANCE trial. Health Qual Life Outcomes. 2007;5:21.

[102] Huang IC, Willke R, Atkinson M, Lenderking W, Frangakis C, Wu A. US and UK versions of the EQ-5D preference weights: Does choice of preference weights make a difference? Quality of Life Research. 2007;16(6):1065-72.

[103] Dolan P. Modeling Valuations for EuroQol Health States. Medical Care. 1997 November;35(11):1095-108.

[104] Kind P. The EuroQol instrument: an index of health-related quality of life. In: Spilker B, editor. Quality of Life and Pharmacoeconomics in Clinical Trials. 2nd ed. ed. Philadelphia: Lippincott-Raven; 1996. p. 191–201.

[105] EuroQOL Group. EuroQol--a new facility for the measurement of health-related quality of life. The EuroQol Group. Health Policy. 1990 Dec;16(3):199-208.

[106] Feeny D, Furlong W, Torrance GW, Goldsmith CH, Zhu Z, DePauw S, et al. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. Med Care. 2002 Feb;40(2):113-28.

[107] Brazier J, Usherwood T, Harper R, Thomas K. Deriving a Preference-Based Single Index from the UK SF-36 Health Survey. Journal of Clinical Epidemiology. 1998;51(11):1115-28.

[108] Brazier J, Ratcliffe J, Salomon J, Tsuchiya A. Measuring and Valuing Health Benefits for Economic Evaluation. Oxford: Oxford University Press; 2007.

[109] Feeny D. Commentary on Jack Dowie,‘‘Decision validity should determine whether a generic or condition-specific HRQOL measure is used in health care decisions’’. Health Econ. 2002;11(1):13-6.

[110] Krahn M, Bremner KE, Tomlinson G, Ritvo P, Irvine J, Naglie G. Responsiveness of disease-specific and generic utility instruments in prostate cancer patients. Quality of Life Research. 2007;V16(3):509-22.

[111] McTaggart-Cowan H, Marra C, Yang Y, Brazier J, Kopec J, FitzGerald J, et al. The validity of generic and condition-specific preference-based instruments: the ability to discriminate asthma control status. Quality of Life Research. 2008;17(3):453-62.

[112] Wiebe S, Guyatt G, Weaver B, Matijevic S, Sidwell C. Comparative responsiveness of generic and specific quality-of-life instruments. J Clin Epidemiol. 2003;56(1):52-60.

[113] Yang Y, Brazier J, Tsuchiya A, Coyne K. Estimating a Preference-Based Single Index from the Overactive Bladder Questionnaire. Value in Health (Online Early Articles). 2008 Jul 18;doi 10.1111/j.1524-4733.2008.00413.x.

[114] Joore M, Brunenberg D, Zank H, van der Stel H, Anteunis L, Boas G, et al. Development of a questionnaire to measure hearing-related health state preferences

framed in an overall health perspective. Int J Technol Assess Health Care. 2002 Summer;18(3):528-39.

[115] Pickard AS, Wilke CT, Lin HW, Lloyd A. Health Utilities Using the EQ-5D in Studies of Cancer. Pharmacoeconomics. 2007;25(5):365-84.

[116] Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. Journal of Health Economics. 2002;21(2):271-92.

[117] Rosser RM. A health index and output measure. In: Walker SR, Rosser RM, editors. Quality of life: assessment and application. Lancaster, England: MTP Press; 1987. p. pp133-60.

[118] Gudex C, Kind P. The QALY toolkit: Centre for Health Economics, University of York; 1988.

[119] Stevens K, McCabe C, Brazier J, Roberts J. Multi-attribute utility function or statistical inference models: A comparison of health state valuation models using the HUI2 health state classification system. Journal of Health Economics. 2007 1 September;26(5):992-1002

[120] Keeney RL, Raiffa H. Decisions with multiple objectives: preferences and value tradeoffs. Cambridge [England] ; New York, NY, USA: Cambridge University Press; 1993.

[121] Hawthorne G, Richardson J, Day NA. A comparison of the Assessment of Quality of Life (AQoL) with four other generic utility instruments. Ann Med. 2001 Jul;33(5):358-70.

[122] Yang Y, Tsuchiya A, Brazier J, Young T. Estimating a preference-based single index from the Asthma Quality of Life Questionnaire (AQLQ). University of Sheffield Health Economics and Decision Science Discussion Paper Series No 07/02. 2007.

[123] Stockler MR, Toner GC, Lewis C, Craft PS, Boyer MJ, Gurney H, et al., editors. Health-related quality of life (HRQL) in a randomised trial of two standard chemotherapy regimens for good-prognosis germ cell tumours: the ANZ germ cell trials group good prognosis trial (Abstract 743). Proc Am Soc Clin Oncol; 2002.

[124] Stockler M, Sourjina T, Grimison P, Gebski V, Byrne M, Harvey V, et al. A randomized trial of capecitabine (C) given intermittently (IC) versus continuously (CC) versus classical CMF as first line chemotherapy for advanced breast cancer (ABC). J Clin Oncol, ASCO Annual Meeting Proceedings (Post-Meeting Edition). 2007;25(18S):1031.

[125] Nowak AK, Byth K, Gebski V, Fong AK, Coates AS, Harvey V, et al. What troubles women starting chemotherapy (CT) for advanced breast cancer (ABC) in a randomized trial? (Abstract). Journal of Clinical Oncology. 2004 July 15;22(14S - July 15 Supplement):8144.

[126] Grimison PS, Coates AS, Forbes JF, Cuzick J, Furnival C, Craft PS, et al. Tamoxifen (TAM) for the prevention of breast cancer: importance of specific aspects of health-related quality of life (HRQL) to global health status in the ANZ BCTG substudy of IBIS-1 (ANZ 92P1) (Abstract 1516). J Clin Oncol, ASCO Annual Meeting Proceedings. 2008;25:15S-1516.

[127] Stockler MR, O'Connell R, Nowak AK, Goldstein D, Turner J, Wilcken NR, et al. Effect of sertraline on symptoms and survival in patients with advanced cancer, but without major depression: a placebo-controlled double-blind randomised trial. Lancet Oncol. 2007 Jun 2.

[128] Martin AJ, Glasziou PP, Simes RJ. A cardiovascular extension of the Health Measurement Questionnaire. J Epidemiol Community Health. 1999 Sept 1;53(9):548-57.

[129] Martin AJ, Glasziou PP, Simes RJ, Lumley T. Predicting patients' utilities from quality of life items: an improved scoring system for the UBQ-H. Quality of Life Research. 1998;7(8):703-11.

[130] Coates A, Glasziou P, McNeil D. On the receiving end--III. Measurement of quality of life during cancer chemotherapy. Ann Oncol. 1990;1(3):213-7.

[131] Priestman TJ, Baum M. Evaluation of quality of life in patients receiving treatment for advanced breast cancer. Lancet. 1976 Apr 24;307(7965):899-900.

[132] Ware JE, Jr., Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care. 1992 Jun;30(6):473-83.

[133] Neugarten BL, Havighurst RJ, Tobin SS. The measurement of life satisfaction. J Gerontol. 1961 Apr;16:134-43.

[134] Coates A, Gebski V, Signorini D, Murray P, McNeil D, Byrne M, et al. Prognostic value of quality-of-life scores during chemotherapy for advanced breast cancer. Australian New Zealand Breast Cancer Trials Group. J Clin Oncol. 1992 Dec 1;10(12):1833-8.

[135] Oken MM, Creech RH, Tormey DC, Horton J, Davis TE, McFadden ET, et al. Toxicity and response criteria of the Eastern Cooperative Oncology Group. Am J Clin Oncol. 1982 Dec;5(6):649-55.

[136] International Breast Cancer Study Group. Multicycle Dose-Intensive Chemotherapy for Women With High-Risk Primary Breast Cancer: Results of International Breast Cancer Study Group Trial 15-95. J Clin Oncol. 2006 Jan 20;24(3):370-8.

[137] Bernhard J, Cella DF, Coates AS, Fallowfield L, Ganz PA, Moinpour CM, et al. Missing quality of life data in cancer clinical trials: serious problems and challenges. Stat Med. 1998 Mar 15-Apr 15;17(5-7):517-32.

[138] Armitage P, Berry G, Matthews JNS. Statistical Methods in Medical Research. Massachusetts: Blackwell Publishing; 2002.

[139] Bernheim JL. How to Get Serious Answers to the Serious Question: 'How have you been?': Subjective Quality of Life (QOL) as an Individual Experiential Emergent Construct. Bioethics. 1999;13(3-4):272-87.

[140] Gudex C, Dolan P, Kind P, Williams A. Health state valuations from the general public using the visual analogue scale. Qual Life Res. 1996 Dec;5(6):521-31.

[141] Dolan P, Culyer AJ, Newhouse JP. The measurement of health-related quality of life for use in resource allocation decisions in health care (Chapter 32). Handbook of Health Economics: Elsevier; 2000. p. 1723-60.

[142] Busschbach JJV, McDonnell J, Essink-Bot M-L, van Hout BA. Estimating parametric relationships between health description and health valuation with an application to the EuroQol EQ-5D. Journal of Health Economics. 1999;18(5):551-70.

[143] Wu EQ, Mulani P, Farrell MH, Sleep D. Mapping FACT-P and EORTC QLQ-C30 to patient health status measured by EQ-5D in metastatic hormone-refractory prostate cancer patients. Value Health. 2007 Sep-Oct;10(5):408-14.

[144] Stockler MR, Osoba D, Goodwin P, Corey P, Tannock IF. Responsiveness to Change in Health-Related Quality of Life in a Randomized Clinical Trial: A Comparison of the Prostate Cancer Specific Quality of Life Instrument (PROSQOLI) with Analogous Scales from the EORTC QLQ-C30 and a Trial Specific Module. Journal of Clinical Epidemiology. 1998;51(2):137-45.

[145] SAS Institute Inc. SAS System for Windows Release 8.02. Cary, N.C., USA2001.

[146] Grimison P, Simes RJ, Hudson HM, Stockler MR. Preliminary validation of an optimally-weighted patient-based utility index by application to randomised trials in breast cancer. Value Health. 2009;12(6):967-76.

[147] Revicki DA, Feeny D, Hunt TL, Cole BF. Analyzing Oncology Clinical Trial Data Using the Q-TWiST Method: Clinical Importance and Sources for Health State Preference Data. Quality of Life Research. 2006 April;15(3):411-23.

[148] Yi MS. The Implications of Differing Methods of Utility Assessment for Patient-Specific Decision-Analytic Tools. Med Decis Making. 2004 Oct 1;24(5):536-7.

[149] Wittenberg E, Winer EP, Weeks JC. Patient utilities for advanced cancer: effect of current health on values. Med Care. 2005 Feb;43(2):173-81.

[150] Stockler MR, Grimison PS, Price TJ, van Hazel GA, Robinson BA, Broad A, et al. Comparing utilities for advanced colorectal cancer valued from societal and cancer-patients' perspectives using baseline data from the MAX study. J Clin Onc, ASCO Annual Meeting Proceedings. 2008;25:15S-6504.

[151] Brazier J, Deverill M. A checklist for judging preference-based measures of health related quality of life: learning from psychometrics. Health Econ. 1999 Feb;8(1):41-51.

[152] Kaplan RM, Bush JW, Berry CC. Health status: types of validity and the index of well-being. Health Serv Res. 1976 Winter;11(4):478-507.

[153] Norinder AL, Krabbe PFM, editors. On the relationship between trade-off and scaling techniques for the valuation of health studies. 19th Plenary Meeting of the EuroQol Group; 2002 13th – 14th September; York, United Kingdom.

[154] O'Leary JF, Fairclough DL, Jankowski MK, Weeks JC. Comparison of time-tradeoff utilities and rating scale values of cancer patients and their relatives: evidence for a possible plateau relationship. Med Decis Making. 1995 Apr-Jun;15(2):132-7.

[155] Wolfson AD, Sinclair AJ, Bombardier C, McGeer A. Preference measurement for functional status in stroke patients: inter-rater and inter-technique comparisons. In: Kane RL, Kane RA, editors. Values and long-term care. Lexington, Mass.: Lexington Books; 1982. p. 191-214.

[156] McCabe C, Stevens K, Roberts J, Brazier J. Health state values for the HUI 2 descriptive system: results from a UK survey. Health Economics. 2005;14(3):231-44.

[157] Stevens KJ, McCabe CJ, Brazier JE. Mapping between Visual Analogue Scale and Standard Gamble data; results from the UK Health Utilities Index 2 valuation survey. Health Economics. 2006;15(5):527-33.

[158] Le Galès C, Buron C, Costet N, Rosman S, Slama PG. Development of a Preference-Weighted Health Status Classification System in France: The Health Utilities Index 3. Health Care Management Science. 2002;V5(1):41-51.

[159] Robinson A, Loomes G, Jones-Lee M, Robinson A, Loomes G, Jones-Lee M. Visual analog scales, standard gambles, and relative risk aversion. Medical Decision Making. 2001 Jan-Feb;21(1):17-27.

[160] Krabbe PF, Essink-Bot ML, Bonsel GJ. The comparability and reliability of five health-state valuation methods. Soc Sci Med. 1997 Dec;45(11):1641-52.

[161] Feeny D, Townsend M, Furlong W, Tonmkins D, Robinson G, Torrance G, et al. Assessing Health-Related Quality-of-Life in Prenatal Diagnosis Comparing Chorionic Villi Sampling and Amniocentesis: A Technical Report: McMaster University Centre for Health Economics and Policy Analysis Research2000 May Contract No.: Assessing Health-Related Quality-of-Life in Prenatal Diagnosis Comparing Chorionic Villi Sampling …
DJ Tomkins, GE Robinson, GW Torrance, PT Mohide, Q … - chepa.org
... Cite as: Feeny D, Townsend M, Furlong W, Tomkins ... G, Mohide P, Wang Q. Assessing
Health-related Quality-of-life ... Analysis Research Working Paper 00-04, May 2000.
...
Related Articles - Web Search

[162] Torrance GW, Boyle MH, Horwood SP. Application of multi-attribute utility theory to measure social preferences for health states. Oper Res. 1982 Nov-Dec;30(6):1043-69.

[163] Dyer JS, Sarin RK. Relative Risk Aversion. Management Science. 1982;28(8):875-86.

[164] McCabe CJ, Stevens KJ, Brazier JE. Utility scores for the Health Utilities Index Mark 2: an empirical assessment of alternative mapping functions. Med Care. 2005 Jun;43(6):627-35.

[165] Grimison P, Simes RJ, Hudson HM, Stockler MR. Deriving a patient-based utility index from a cancer-specific quality of life questionnaire. Value Health. 2009;12(5):800-7.

[166] Stockler MR, Osoba D, Corey P, Goodwin PJ, Tannock IF. Convergent Discriminitive, and Predictive Validity of the Prostate Cancer Specific Quality of Life Instrument (PROSQOLI) Assessment and Comparison with Analogous Scales From the EORTC QLQ-C30 and a Trial-Specific Module. Journal of Clinical Epidemiology. 1999;52(7):653-66.

[167] Bottomley A, Efficace F. Predicting survival in advanced cancer patients: is it possible with patient-reported health status data? Ann Oncol. 2006 Jul;17(7):1037-8.

[168] Coates AS, Hurny C, Peterson HF, Bernhard J, Castiglione-Gertsch M, Gelber RD, et al. Quality-of-Life Scores Predict Outcome in Metastatic but Not Early Breast Cancer. J Clin Oncol. 2000 Nov 15;18(22):3768-74.

[169] Dancey J, Zee B, Osoba D, Whitehead M, Lu F, Kaizer L, et al. Quality of life scores: an independent prognostic variable in a general population of cancer patients receiving chemotherapy. The National Cancer Institute of Canada Clinical Trials Group. Qual Life Res. 1997 Mar;6(2):151-8.

[170] Siddiqui F, Kachnic LA, Movsas B. Quality-of-life outcomes in oncology. Hematology/oncology clinics of North America. 2006 Feb;20(1):165-85.

[171] Gotay CC, Kawamoto CT, Bottomley A, Efficace F. The Prognostic Significance of Patient-Reported Outcomes in Cancer Clinical Trials. J Clin Oncol. 2008 March 10;26(8):1355-63.

[172] McKenna SP, Ratcliffe J, Meads DM, Brazier JE. Development and validation of a preference based measure derived from the Cambridge Pulmonary

Hypertension Outcome Review (CAMPHOR) for use in cost utility analyses. Health and Quality of Life Outcomes. 2008;6:65.

[173] Glasziou PP, Simes RJ, Gelber RD. Quality adjusted survival analysis. Stat Med. 1990 Nov;9(11):1259-76.

[174] Haneuse S, Wakefield J, Sheppard L. The interpretation of exposure effect estimates in chronic air pollution studies. Statistics in Medicine. 2007;26(16):3172-87.

[175] Dominici F, Sheppard L, Clyde M. Health Effects of Air Pollution: A Statistical Review. International Statistical Review. 2003;71(2):243-76.

[176] 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 Jan;41(2):280-7.

[177] Zeger SL, Liang KY, Albert PS. Models for longitudinal data: a generalized estimating equation approach. Biometrics. 1988 Dec;44(4):1049-60.

[178] Senn S, Stevens L, Chaturvedi N. Repeated measures in clinical trials: simple strategies for analysis using summary measures. Statistics in Medicine. 2000;19(6):861-77.

[179] Fairclough DL. Design and analysis of quality of life studies in clinical trials. Boca Raton, Fla: Chapman & Hall; 2002.

[180] Perez DJ, Williams SM, Christensen EA, McGee RO, Campbell AV. A longitudinal study of health related quality of life and utility measures in patients with advanced breast cancer. Qual Life Res. 2001;10(7):587-93.

[181] Lidgren M, Wilking N, Jönsson B, Rehnberg C. Health related quality of life in different states of breast cancer. Quality of Life Research. 2007;16(6):1073-81.

[182] Lloyd A, Nafees B, Narewska J, Dewilde S, Watkins J. Health state utilities for metastatic breast cancer. Br J Cancer. 2006 Sep 18;95(6):683-90.

[183] Dranitsaris G, Leung P, Mather J, Oza A. Cost-utility analysis of second-line hormonal therapy in advanced breast cancer: a comparison of two aromatase inhibitors to megestrol acetate. Anti-cancer drugs. 2000 Aug;11(7):591-601.

[184] Dranitsaris G, Verma S, Trudeau M. Cost utility analysis of first-line hormonal therapy in advanced breast cancer: comparison of two aromatase inhibitors to tamoxifen. Am J Clin Oncol. 2003 Jun;26(3):289-96.

[185] Marchetti M, Caruggi M, Colombo G. Cost utility and budget impact of third-generation aromatase inhibitors for advanced breast cancer: a literature-based model analysis of costs in the Italian national health service. Clinical Therapeutics. 2004;26(9):1546-61.

[186] Grann VR, Jacobson JS, Sundararajan V, Albert SM, Troxel AB, Neugut AI. The quality of life associated with prophylactic treatments for women with BRCA1/2 mutations. The cancer journal from Scientific American. 1999 Sep-Oct;5(5):283-92.

[187] Grann VR, Sundararajan V, Jacobson JS, Whang W, Heitjan DF, Antman KH, et al. Decision analysis of tamoxifen for the prevention of invasive breast cancer. Cancer J (Sudbury, Mass). 2000 May-Jun;6(3):169-78.

[188] Hershman D, Sundararajan V, Jacobson JS, Heitjan DF, Neugut AI, Grann VR. Outcomes of Tamoxifen Chemoprevention for Breast Cancer in Very High-Risk Women: A Cost-Effectiveness Analysis. J Clin Oncol. 2002 Jan 1;20(1):9-16.

[189] Milne RJ, Heaton-Brown KH, Hansen P, Thomas D, Harvey V, Cubitt A. Quality-of-life valuations of advanced breast cancer by New Zealand women. Pharmacoeconomics. 2006;24(3):281-92.

[190] de Cock E, Hutton J, Canney P, Body J, Barrett-Lee P, Neary M, et al. Cost-effectiveness of oral ibandronate compared with intravenous (i.v.) zoledronic acid or

i.v. generic pamidronate in breast cancer patients with metastatic bone disease undergoing i.v. chemotherapy. Supportive Care in Cancer. 2005;13(12):975-86.

[191] de Cock E, Hutton J, Canney P, Body JJ, Barrett-Lee P, Neary MP, et al. Cost-effectiveness of oral ibandronate versus IV zoledronic acid or IV pamidronate for bone metastases in patients receiving oral hormonal therapy for breast cancer in the United Kingdom. Clinical Therapeutics. 2005;27(8):1295-310.

[192] van den Hout WB, van der Linden YM, Steenland E, Wiggenraad RGJ, Kievit J, de Haes H, et al. Single- Versus Multiple-Fraction Radiotherapy in Patients With Painful Bone Metastases: Cost-Utility Analysis Based on a Randomized Trial. J Natl Cancer Inst. 2003 Feb 5;95(3):222-9.

[193] Cykert S, Phifer N, Hansen C. Tamoxifen for breast cancer prevention: a framework for clinical decisions. Obstetrics and gynecology. 2004 Sep;104(3):433-42.

[194] Hillner BE, Weeks JC, Desch CE, Smith TJ. Pamidronate in Prevention of Bone Complications in Metastatic Breast Cancer: A Cost-Effectiveness Analysis. J Clin Oncol. 2000 Jan 1;18(1):72-9.

[195] Launois R, Reboul-Marty J, Henry B, Bonneterre J. A cost-utility analysis of second-line chemotherapy in metastatic breast cancer. Docetaxel versus paclitaxel versus vinorelbine. Pharmacoeconomics. 1996 Nov;10(5):504-21.

[196] Verma S, Maraninchi D, O'Shaughnessy J, Jamieson C, Jones S, Martin M, et al. Capecitabine plus docetaxel combination therapy. Cancer. 2005 Jun 15;103(12):2455-65.

[197] Brown RE, Hutton J. Cost-utility model comparing docetaxel and paclitaxel in advanced breast cancer patients. Anti-cancer drugs. 1998 Nov;9(10):899-907.

[198] Brown RE, Hutton J, Burrell A. Cost Effectiveness of Treatment Options in Advanced Breast Cancer in the UK. Pharmacoeconomics. 2001 Nov;19(11):1091-102.

[199] Martin SC, Gagnon DD, Zhang L, Bokemeyer C, Van Marwijk Kooy M, van Hout B. Cost-utility analysis of survival with epoetin-alfa versus placebo in stage IV breast cancer. Pharmacoeconomics. 2003;21(16):1153-69.

[200] Elkin EB, Weinstein MC, Winer EP, Kuntz KM, Schnitt SJ, Weeks JC. HER-2 Testing and Trastuzumab Therapy for Metastatic Breast Cancer: A Cost-Effectiveness Analysis. J Clin Oncol. 2004 Mar 1;22(5):854-63.

[201] Hornberger J, Cosler LE, Lyman GH. Economic analysis of targeting chemotherapy using a 21-gene RT-PCR assay in lymph-node-negative, estrogen-receptor-positive, early-stage breast cancer. The American journal of managed care. 2005 May;11(5):313-24.

[202] Hillner BE, Smith TJ. Efficacy and cost effectiveness of adjuvant chemotherapy in women with node-negative breast cancer. A decision-analysis model. N Engl J Med. 1991 Jan 17;324(3):160-8.

[203] Hillner BE, Smith TJ, Desch CE. Efficacy and cost-effectiveness of autologous bone marrow transplantation in metastatic breast cancer. Estimates using decision analysis while awaiting clinical trial results. Jama. 1992 Apr 15;267(15):2055-61.

[204] Hillner BE. Benefit and projected cost-effectiveness of anastrozole versus tamoxifen as initial adjuvant therapy for patients with early-stage estrogen receptor-positive breast cancer. Cancer. 2004 Sep 15;101(6):1311-22.

[205] Hayman JA, Hillner BE, Harris JR, Weeks JC. Cost-effectiveness of routine radiation therapy following conservative surgery for early-stage breast cancer. J Clin Oncol. 1998 Mar 1;16(3):1022-9.

[206] Hutton J, Brown R, Borowitz M, Abrams K, Rothman M, Shakespeare A. A new decision model for cost-utility comparisons of chemotherapy in recurrent metastatic breast cancer. Pharmacoeconomics. 1996;9 Suppl 2:8-22.

[207] Lonning PE. Comparing cost/utility of giving an aromatase inhibitor as monotherapy for 5 years versus sequential administration following 2-3 or 5 years of tamoxifen as adjuvant treatment for postmenopausal breast cancer. Ann Oncol. 2006 Feb;17(2):217-25.

[208] Scuffham PA, Whitty JA, Mitchell A, Viney R. The use of QALY weights for QALY calculations: a review of industry submissions requesting listing on the Australian Pharmaceutical Benefits Scheme 2002-4. Pharmacoeconomics. 2008;26(4):297-310.

[209] De Wit GA, Busschbach JJV, De Charro FT. Sensitivity and perspective in the valuation of health status: whose values count? Health Economics. 2000;9(2):109-26.

[210] Buckingham K, Devlin NJ, Hansen P. Does it matter whose valuations are used to estimate health state tariffs, and which tariffs are used for CUA? 17th Plenary Meeting of the Euroqol Group; Spain2000.

[211] Chapman RH, Berger M, Weinstein MC, Weeks JC, Goldie S, Neumann PJ. When does quality-adjusting life-years matter in cost-effectiveness analysis? Health Econ. 2004 May;13(5):429-36.

Appendices

Appendix 1 Utility-Based Questionnaire-Cancer (UBQ-C)

UBQ-C Questionnaire (Page 1 of 3)

Randomisation number: ☐☐☐☐☐☐☐ Patient's initials ☐☐☐

Today's Date: ☐☐/☐☐/☐☐☐☐ Institution/Hospital:_____

Place of completion (☑): **Clinic ☐ Home ☐**

1. How would you describe your **general health** over the **past 3-4 weeks?**
*Please tick (☑) the **one** most appropriate answer.*

Excellent ..... ☐
Very good ..... ☐
Good ..... ☐
Fair ..... ☐
Poor ..... ☐

2. Has your health affected your ability to perform any of the following **activities** over the **past 3-4 weeks?** *Please tick (☑) the **one** most appropriate answer for each activity.*

	Not affected	Slightly affected	Severely activities	Unable to do activities at all
Usual daily activities (eg. paid work, house chores, etc.)...	☐	☐	☐	☐
Your social life.....	☐	☐	☐	☐
Your hobbies or leisure activities.....	☐	☐	☐	☐
Your sex life	☐	☐	☐	☐
Walking several blocks (500 metres)	☐	☐	☐	☐
Climbing one flight of stairs	☐	☐	☐	☐
Vigorous activities such as running or strenuous sports	☐	☐	☐	☐
Self-care (washing, dressing etc.)	☐	☐	☐	☐

UBQ-C Questionnaire (Page 2 of 3)

Randomisation number: ☐☐☐☐☐☐☐ Patient's initials ☐☐☐

Today's Date: ☐☐/☐☐/☐☐☐☐ Institution/Hospital:_____

Place of completion (☑): **Clinic ☐ Home ☐**

3. Over the past 3-4 weeks have you experienced any of the following problems? If no, please circle (NONE). If yes, circle on a scale from 1 to 10 how much distress this has caused you.

	AMOUNT OF DISTRESS										
	NONE		**MILD**		**MODERATE**			**SEVERE**			**EXTREME**
Shortness of breath	0	1	2	3	4	5	6	7	8	9	10
Difficulty sleeping	0	1	2	3	4	5	6	7	8	9	10
Feeling sick (nausea/vomiting)	0	1	2	3	4	5	6	7	8	9	10
Lack of energy	0	1	2	3	4	5	6	7	8	9	10
Aches and pain	0	1	2	3	4	5	6	7	8	9	10
Feeling sad or depressed	0	1	2	3	4	5	6	7	8	9	10
Feeling anxious or worried	0	1	2	3	4	5	6	7	8	9	10
Loss of appetite	0	1	2	3	4	5	6	7	8	9	10
Dissatisfaction with your weight or appearance	0	1	2	3	4	5	6	7	8	9	10
Uncertainty about the future	0	1	2	3	4	5	6	7	8	9	10
Numbness or pins & needles	0	1	2	3	4	5	6	7	8	9	10
Anger or resentment	0	1	2	3	4	5	6	7	8	9	10
Loneliness	0	1	2	3	4	5	6	7	8	9	10
Loss of hair	0	1	2	3	4	5	6	7	8	9	10
Diarrhoea	0	1	2	3	4	5	6	7	8	9	10
Constipation	0	1	2	3	4	5	6	7	8	9	10
Loss of self confidence	0	1	2	3	4	5	6	7	8	9	10
Feeling dependent on others	0	1	2	3	4	5	6	7	8	9	10
Thought of chemotherapy	0	1	2	3	4	5	6	7	8	9	10
Inability to concentrate	0	1	2	3	4	5	6	7	8	9	10
Any other problems that cause you distress *(if none circle '0')*	0	1	2	3	4	5	6	7	8	9	10

Please list: _____

UBQ-C Questionnaire (Page 3 of 3)

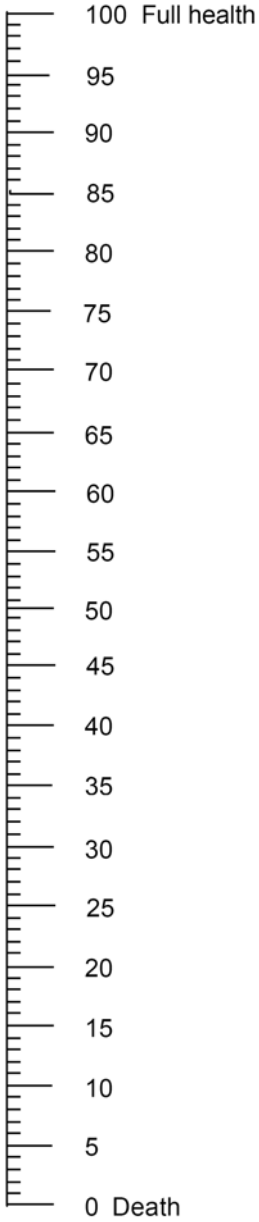
Randomisation number: ☐☐☐☐☐☐☐ Patient's initials ☐☐☐

Today's Date: ☐☐/☐☐/☐☐☐☐ Institution/Hospital:_____

Place of completion (☑): **Clinic ☐ Home ☐**

4. To help people say exactly how good or bad their health is, we have drawn a scale (rather like a thermometer) on which full health is marked by 100 and death is marked by 0. We would like you to indicate on this scale how good or bad your own health is today, in your opinion. Please do this by drawing a line from the box below to whichever point on the scale indicates how good or bad your current health state is.

Your own health state today



Appendix 2 Other questionnaires

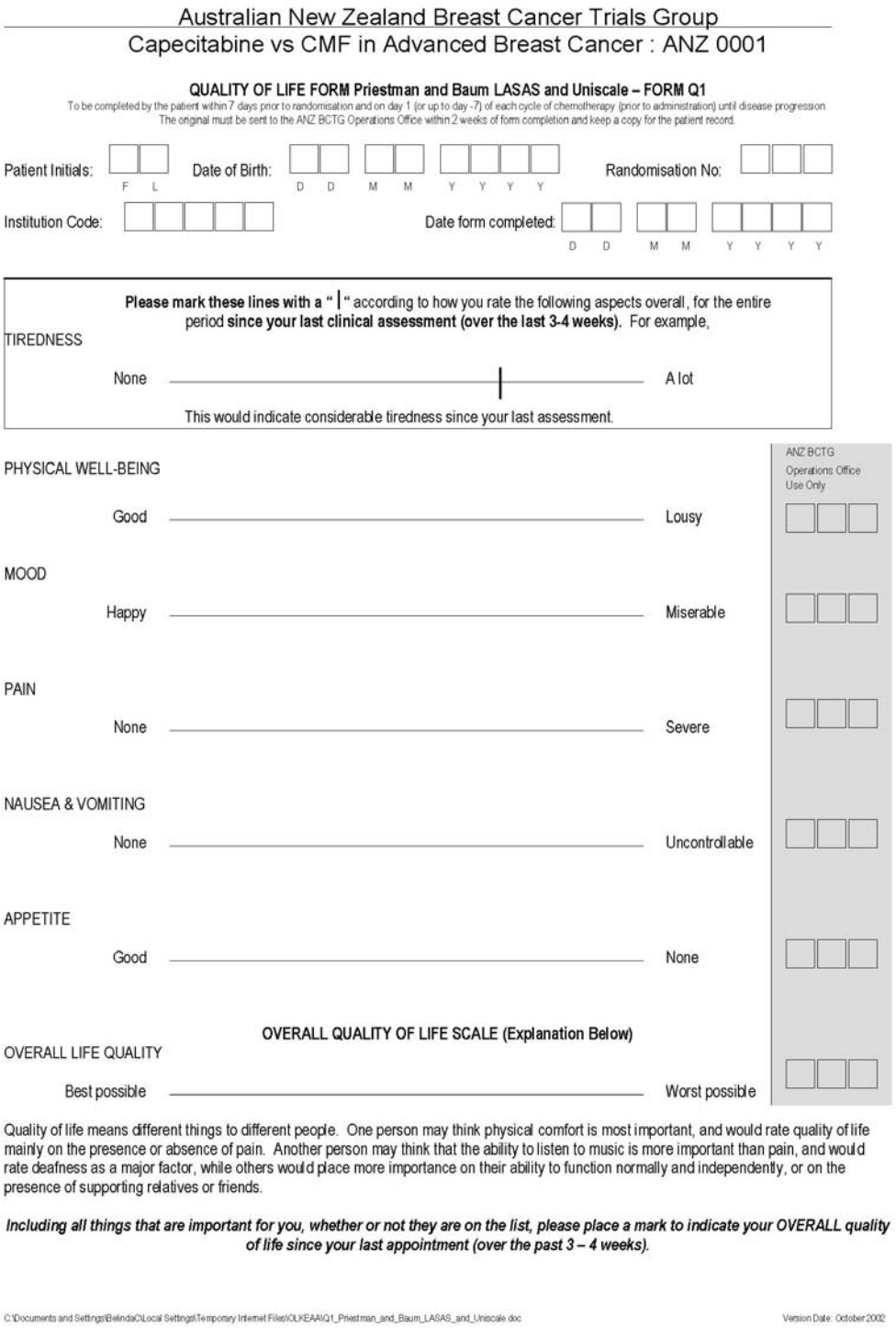
1. Priestman and Baum Linear Analog Self Assessment Scales (LASAS)
2. Spitzer-Uniscale
3. Chemotherapy Acceptability Questionnaire
4. Eastern Cooperative Oncology Group (ECOG) Performance Status Scale

Priestman and Baum Linear Analog Self Assessment Scales (LASAS):

1st five scales

Spitzer-Uniscale:

Labelled 'Overall life quality'



Australian New Zealand Breast Cancer Trials Group

Capecitabine vs CMF in Advanced Breast Cancer : ANZ 0001

QUALITY OF LIFE FORM Priestman and Baum LASAS and Uniscale – FORM Q1

To be completed by the patient within 7 days prior to randomisation and on day 1 (or up to day -7) of each cycle of chemotherapy (prior to administration) until disease progression. The original must be sent to the ANZ BCTG Operations Office within 2 weeks of form completion and keep a copy for the patient record.

Patient Initials: F L Date of Birth: D D M M Y Y Y Y Randomisation No:

Institution Code: Date form completed: D D M M Y Y Y Y

Please mark these lines with a " | " according to how you rate the following aspects overall, for the entire period since your last clinical assessment (over the last 3-4 weeks). For example,

TIREDNESS

None ——|—— A lot

This would indicate considerable tiredness since your last assessment.

ANZ BCTG Operations Office Use Only

PHYSICAL WELL-BEING

Good —— Lousy

MOOD

Happy —— Miserable

PAIN

None —— Severe

NAUSEA & VOMITING

None —— Uncontrollable

APPETITE

Good —— None

OVERALL QUALITY OF LIFE SCALE (Explanation Below)

OVERALL LIFE QUALITY

Best possible —— Worst possible

Quality of life means different things to different people. One person may think physical comfort is most important, and would rate quality of life mainly on the presence or absence of pain. Another person may think that the ability to listen to music is more important than pain, and would rate deafness as a major factor, while others would place more importance on their ability to function normally and independently, or on the presence of supporting relatives or friends.

Including all things that are important for you, whether or not they are on the list, please place a mark to indicate your OVERALL quality of life since your last appointment (over the past 3 – 4 weeks).

C:\Documents and Settings\BelindaC\Local Settings\Temporary Internet Files\OLKEAA\Q1_Priestman_and_Baum_LASAS_and_Uniscale.doc Version Date: October 2002

Chemotherapy Acceptability Questionnaire:

Australian New Zealand Breast Cancer Trials Group

Capecitabine vs CMF in Advanced Breast Cancer : ANZ 0001

QUALITY OF LIFE FORM Inconvenience and side effects of therapy – FORM Q3

To be completed by the patient within 7 days prior to randomisation and on day 1 (or up to day -7) of each cycle of chemotherapy (prior to administration) until disease progression.
The original must be sent to the ANZ BCTG Operations Office within 2 weeks of form completion and a copy kept for the patient record.

Patient Initials: ☐☐ (F L) Date of Birth: ☐☐ (D D) ☐☐ (M M) ☐☐☐☐ (Y Y Y Y) Randomisation No: ☐☐☐

Institution Code: ☐☐☐☐☐ Date form completed: ☐☐ (D D) ☐☐ (M M) ☐☐☐☐ (Y Y Y Y)

Please circle one number for each line to show how much that aspect has troubled you on average over the last 3-4 weeks.

	No trouble at all		Mild			Moderate			Severe		Worst I can imagine
1. Sore mouth	0	1	2	3	4	5	6	7	8	9	10
2. Altered taste	0	1	2	3	4	5	6	7	8	9	10
3. Sore or watery eyes	0	1	2	3	4	5	6	7	8	9	10
4. Nausea	0	1	2	3	4	5	6	7	8	9	10
5. Vomiting	0	1	2	3	4	5	6	7	8	9	10
6. Skin rash	0	1	2	3	4	5	6	7	8	9	10
7. Sore hands or feet	0	1	2	3	4	5	6	7	8	9	10
8. Difficulty walking	0	1	2	3	4	5	6	7	8	9	10
9. Problems taking tablets	0	1	2	3	4	5	6	7	8	9	10
10. Problems with needles or injections	0	1	2	3	4	5	6	7	8	9	10
11. Inconvenience of treatment	0	1	2	3	4	5	6	7	8	9	10
12 Problems coping with treatment	0	1	2	3	4	5	6	7	8	9	10
13. Problems looking after myself	0	1	2	3	4	5	6	7	8	9	10
14. Problems doing what I want	0	1	2	3	4	5	6	7	8	9	10
15. Problems for my friends & family	0	1	2	3	4	5	6	7	8	9	10

Thank you for completing this form.

K:\Trials\Capecitabine\Forms\CRFs October 2002\Q3_Inconvenience_and_side_effects_of_therapy.doc Version Date: October2002

ECOG Performance Status Scale

Grade	ECOG
0	Fully active, able to carry on all pre-disease performance without restriction
1	Restricted in physically strenuous activity but ambulatory and able to carry out work of a light or sedentary nature, e.g. light house work, office work
2	Ambulatory and capable of all self-care but unable to carry out any work activities. Up and about more than 50% of waking hours
3	Capable of only limited self-care, confined to bed or chair more than 50% of waking hours
4	Completely disabled. Cannot carry on any self-care. Totally confined to bed or chair

As published in Am. J. Clin. Oncol.:
Oken, M.M., Creech, R.H., Tormey, D.C., Horton, J., Davis, T.E., McFadden, E.T., Carbone, P.P.: Toxicity And Response Criteria Of The Eastern Cooperative Oncology Group. Am J Clin Oncol 5:649-655, 1982

Appendix 3 Patient information sheet and consent form for valuation survey

RESEARCH STUDY INTO THE DEVELOPMENT OF A UTILITY BASED QUALITY OF LIFE QUESTIONNAIRE IN PATIENTS WITH CANCER

INFORMATION FOR PARTICIPANTS

You are invited to take part in a research study into the development of a Quality of Life Questionnaire in patients with cancer. The objective is to develop a short health questionnaire for cancer patients. Such a questionnaire could be used to monitor patients well-being over time and could be used to evaluate new treatments. The study is being conducted by Professor John Simes, Elaine Beller, Andrew Martin, of the Clinical Trials Centre part of the National Health and Medical Research Council, in association with the New South Wales Cancer Council.

If you agree to participate in this study, you will be given a short questionnaire which should take you about 15 minutes to complete. Shortly after you will receive another questionnaire of a similar length. After completing the second questionnaire you will be asked to take part in an interview at your outpatient clinic at a time which is convenient for you. In the interview you will be presented with descriptions of different health conditions and you will be asked to choose which you think sounds best. Although no physical risks are involved it is possible that some people may be upset by interview questions relating to their illness.

All aspects of the study, including results, will be strictly confidential and not revealed to any medical staff looking after you. A report of the study may be submitted for publication, but individual participants will not be identifiable in such a report.

While we intend that this research study furthers medical knowledge and may help identify better treatments, it may not be of direct benefit to you.

Participation in this study is entirely voluntary: you are not obliged to participate and - if you do participate - you can withdraw at any time. Whatever your decision it will not affect your medical treatment or your relationship with medical staff.

When you have read this information, Vicki Greatorex, will discuss it with you further and answer any questions you may have. If you would like to know more at any stage, please feel free to contact Vicki Greatorex (Research Assistant) on (02) 692 4561. This information sheet is for you to keep.

Any person with concerns or complaints about the conduct of a research study can contact the Secretary of the Ethics Review Committee of the Central Sydney Area Health Service on (02) 516 6766.

RESEARCH STUDY INTO THE DEVELOPMENT OF A UTILITY BASED QUALITY OF LIFE QUESTIONNAIRE IN PATIENTS WITH CANCER

PARTICIPANT CONSENT FORM

I, …… *[name]* of

……*[address]* have read and understood the Information for Participants on the above named research study and have discussed the study with ……

I am aware of the procedures involved in the study, including any inconvenience, risk, discomfort or side effect, and of their implications.

I freely choose to participate in this study and understand that I can withdraw at any time.

I also understand that the research study is strictly confidential.

I hereby agree to participate in this research study.

NAME: ……

SIGNATURE: ……

DATE: ……

NAME OF WITNESS: ……

SIGNATURE OF WITNESS: ……

Appendix 4 Patient information sheet and consent form for advanced breast cancer trial

INFORMATION FOR PARTICIPANTS & SAMPLE CONSENT FORM

To be printed on the local institutions letterhead

ANZ BCTG Protocol 0001

Capecitabine vs CMF in Advanced Breast Cancer

A phase III trial to evaluate oral chemotherapy with capecitabine versus standard chemotherapy with CMF in advanced breast cancer.

Principal Investigator: **Principal Investigator Name**

Associate Investigators: **Co-investigator Name**

Co-investigator Name

INFORMATION FOR PARTICIPANTS

You are invited to take part in a research project. The next few pages describe the project in detail. Its purpose is to explain to you as openly and clearly as possible what being in this project involves, before you decide whether to take part.

Please read this Participant Information carefully. Feel free to ask questions about any information in the document. You may also wish to discuss the project with a relative or friend or your local health worker. Feel free to do this.

Once you understand what the project is about, and if you agree to take part in it, you will be asked to sign the Consent Form. By signing the Consent Form, you indicate that you understand the information and that you give your consent to participate in the research project.

You will be given a copy of the Participant Information and Consent Form to keep as a record.

The purpose of this study is to compare a new form of chemotherapy given daily by mouth (capecitabine) to the standard form of chemotherapy given as a combination of tablets and injections given intermittently (CMF). The study also tests whether it is better to give capecitabine daily over 21 days, or to give the same dose spread over 14 out of every 21 days.

Who is conducting the study?

This study is being carried out in hospitals around Australia and New Zealand, and is being coordinated by the Australian and New Zealand Breast Cancer Trials Group. At this hospital, the study is being carried out by Principal Investigator and colleagues from the Department, Centre. The pharmaceutical company Roche is supplying some of the study medication.

How many people will take part?

465 people like you will take part in this study.

What is involved?

If you are interested in taking part in this study, then some tests need to be done to make sure that it is suitable for you. These will include:

- a full medical history (it is important that we know if you have any other medical conditions or are taking any medications)
- a physical examination by your doctor
- routine blood tests (about one tablespoon)
- imaging tests (for example X-rays, CT scans, and a bone scan) to check the extent of your disease

While receiving treatment, you will have check-ups with your doctor and have blood taken for routine blood tests (about one tablespoon) every 3 to 4 weeks. Scans and x-rays will be done every 12 weeks to check how your disease is responding to the treatment. You will also be asked to fill in forms with simple questions about how you are feeling and getting on ('quality of life questionnaires').

If you agree to take part in this study you will get one of the three treatments described below. The treatment you get will be chosen by a computer using a process known as randomisation. This means that your treatment will be selected at random (like drawing a card from a pack) and that you have an equal chance of getting any one of the three treatments. This is necessary to make sure that the people getting each treatment are comparable. Neither you nor your doctor can choose which treatment you will receive. In this study, everyone gets an active treatment – there is NO placebo group in this study.

Treatments

If you take part in this study you will get one of the three following treatments:

1. CMF (cyclophosphamide, methotrexate, fluorouracil)

2 drugs (M&F) are given by injection into a vein on the 1st and 8th day of each 28 day treatment cycle, and 1 drug (C) is given as pills for the first 14 days of each 28 day treatment cycle. The treatment is given as an outpatient without need for hospital admission and is repeated every 28 days if there are no troublesome side effects.

2. Intermittent capecitabine

Capecitabine is given as pills taken by mouth morning and night for the first 14 days. There are no injections. The treatment is given as an outpatient without need for hospital admission and is repeated every 21 days if there are no troublesome side effects.

3. Continuous capecitabine

Capecitabine is given as pills taken by mouth morning and night for 21 days. There are no injections. The treatment is given as an outpatient without need for hospital admission and is repeated every 21 days if there are no troublesome side effects.

How long will I be in the study?

The chemotherapy will continue so long as it seems to be working, it is not causing troublesome side effects, and both you and your doctor think it should continue. On average, this means treatment for about 6 months, but it may be for as little as a few weeks or as much as a few years. You can choose to stop the treatment at any time without penalty. Once the study treatment is stopped, the choice of further treatment is up to you and your doctor. If you have further treatment we will continue to follow your progress, but you will not need to have any extra visits or tests and you will not be asked to complete any more forms.

What are the possible problems or side effects?

Cancer treatments are often associated with unwanted side effects. You may experience none, some, or all of the effects listed below, and they maybe of mild, moderate or severe intensity. In addition there is always the risk of a previously unknown side effect occurring. If a severe side effect or reaction occurs, your doctor may need to stop your treatment. You should inform your doctor of any problems. In addition to the side effects of treatment, there is a small risk of discomfort or bruising when having blood and imaging tests.

Problems or side effects with CMF

Treatment with CMF includes pills taken by mouth and injections. Some people find it difficult to swallow pills. The injections require placement of a thin tube into a vein (cannula or 'drip') on the 1st and 8th day of each 28 day treatment cycle. Placement of the cannula usually causes a little discomfort but usually no serious problems. It may cause bleeding, bruising, discomfort, pain and rarely infection at or near the insertion point. The possible side effects of CMF include fatigue; altered taste and appetite; nausea and/or vomiting; sore mouth, eyes or abdomen; diarrhoea; hair loss; skin rashes and nail changes; irritation of the bladder; decrease in blood cell counts with the possibility of fever, infection and bleeding. These side effects generally disappear once treatment is stopped.

Problems or side effects with capecitabine

Treatment with capecitabine includes only pills taken by mouth, there are no injections. Some people find it difficult to swallow pills. The possible side effects of capecitabine include fatigue; altered taste and appetite; nausea and/or vomiting; sore mouth, eyes or abdomen; diarrhoea; pins and needles, swelling or rashes of the hands and/or feet; skin rashes and nail changes; hair loss. Decreases in blood cell counts are less common with capecitabine than with CMF. These side effects generally disappear once treatment is stopped.

Other less common problems or side effects with CMF or Capecitabine

Angina (chest pain) and deep vein thrombosis (blood clots in the legs) have been reported in people having CMF, capecitabine and other types of chemotherapy. These problems are uncommon (less than 1% of people having chemotherapy) and are more likely to affect those who have had them before. These problems can be severe, but usually improve with standard treatments.

Pregnancy and contraception

The effects of chemotherapy drugs on the unborn child and on the newborn baby are not known. Because of this, it is important that the study participants are not pregnant or breast feeding and do not become pregnant during the course of the study. You must not participate in the study if you are pregnant or trying to become pregnant, or breast feeding.

If you are female and child bearing is a possibility, you will be required to undergo a pregnancy test prior to commencing the study. Female participants are strongly advised to use effective contraception during the course of the study. If you do become pregnant whilst participating in the study, you should advise your treating doctor immediately. He/she will withdraw you from the study and advise you on further medical attention, should this be necessary. You must not continue in the study if you become pregnant.

Other treatment whilst on study

It is important that you tell your doctor about any treatments or medications you start, stop or change while you are taking part in the study. This includes non-prescription medications, vitamins and herbal remedies.

Do I have to take part in this study?

Participation in this study is voluntary -- you do not have to take part, and if you do take part, you can decide to stop the study treatment at any time. Whatever your decision, it will not affect your treatment or your relationship with the medical or other staff.

If you decide to stop the trial treatment or wish to withdraw from participating in the study, please first notify a member of the research team. This notice will allow that person or the research supervisor to inform you if there are any health risks or special requirements linked to withdrawing. We would also like, with your permission, to continue collecting information about your health status.

What happens if I don't participate?

If you do not wish to participate in this study, there are other treatments available to you. Your doctor will discuss other treatment options available to patients with your type of cancer and explain the risks and benefits of these treatments to you.

Are there any benefits to taking part in this study?

The aim of this research study is to improve medical knowledge and improve the treatment of breast cancer in the future, however taking part in the trial may not be of direct benefit to you. All of the drugs in the study are known to work in breast cancer, and each has potential advantages and disadvantages. It is not clear which is the best treatment.

How will information about me be kept private?

Any information obtained in connection with this study and that can identify you will be stored in strict confidence, as required by law, on computer disk and/or paper file in locked offices in the hospital Medical Oncology Department. Only staff associated with this trial will have access to it and it will only be disclosed with your permission, except as required by law. Once the study is completed, records will be retained in a locked storage facility indefinitely. However, your medical records and any information obtained during the study are subject to inspection (for the purpose of verifying the procedures and the data) by the Australian Government's Therapeutic Goods Administration (TGA) or a New Zealand equivalent, other national drug regulatory authorities (where this is applicable) such as the Food and Drug Administration (FDA) of the United States of America, authorised representatives of ANZ BCTG, and of the pharmaceutical company supplying or distributing the study drugs, Roche Products Pty Ltd and subcontractors. This will be done only under the formal agreement that confidentiality will be respected in all cases.

By signing the consent form, you authorise release of, or access to, this confidential information to the relevant study personnel and regulatory authorities, as noted above.

Data on the treatment you receive, tests done and your progress will be reported in confidence to the ANZ BCTG, who will put this together with data from other patients in the study, without identifying anyone individually. Data about you will not have your name on it. It will have a unique patient identification number that has been assigned when you entered the study.

Under Australian Privacy and other relevant laws, you have the right to access information that is collected and stored about you. You should contact the persons named in the 'Who can I call if I have questions or problems?' section of this document, if you wish to access your information.

What if new information comes to light during the study?

You will be informed of any significant new findings about the trial treatments which occur during the study and which may lead you to change your willingness to participate.

What will happen with the results of the study?

It will take several years for this study to be finished and for its results to become available. The results will be published in medical journals that are available to the

public. If you would like to see these reports, then you should ask your doctor. No report will include information that allows you to be identified.

What are the costs of the study?

Chemotherapy and other medications will be prescribed and paid for as usual in this cancer centre outside of this study. You may have to pay a small amount for some prescriptions (approximately $3.60 for pensioners, $22.40 for others). You will not be paid for taking part in this study.

In the unlikely event of a physical injury as a result of your participation in this study, the parties involved in this study agree to be bound by the Guidelines for Compensation for Injury Resulting from Participation in an Industry-Sponsored Clinical Trial of the Medicines Australia or New Zealand equivalent. A copy of these guidelines is available from the Secretary of the Ethics Review Committee.

The cancer centre responsible for your treatment will receive some payment from the ANZ Breast Cancer Trials Group to offset the costs of running the trial.

Who can I call if I have questions or problems?

When you have read this information, Dr_____ will discuss it with you further and answer any questions you may have. If you would like to know more about the study or treatment, please contact Principal investigator on phone number. This information sheet is for you to keep.

This study has been approved by the Ethics Review Committee – Centre. If you have any concerns or complaints about the conduct of this study, you may contact the Secretary of the Ethics Review Committee – Centre on phone number. Alternatively, if you wish to speak with an independent person within the Hospital about any problems or queries about the way in which the study was conducted, you may contact the Patient Representative on phone number.

ANZ BCTG Protocol 0001
Capecitabine vs CMF in Advanced Breast Cancer

A phase III trial to evaluate oral chemotherapy with capecitabine versus standard chemotherapy with CMF in advanced breast cancer.

Principal Investigator:	**Principal Investigator Name**
Associate Investigators:	**Co-investigator Name**
	Co-investigator Name

PARTICIPANT CONSENT FORM

I, ... *[name]*

Of.....……[a*ddress*]
have read and understood the Information for Participants on the above named research study and have discussed the study with .....

I have been made aware of the procedures involved in the study, including any known or expected inconvenience, risk, discomfort or side effect, and of their implications as far as they are currently known by the researchers.

I understand that the research project will be carried out according to the principles in the National Health & Medical Research Council National Statement on Ethical Conduct in Research Involving Humans (June 1999).

I freely choose to participate in this study and understand that I can withdraw at any time.

I also understand that the research study is strictly confidential.

I hereby agree to participate in this research study.

NAME: .....…

SIGNATURE: .....… **DATE:** .....…

NAME OF WITNESS: .....…

SIGNATURE OF WITNESS: .....… **DATE:** .....

NAME OF INVESTIGATOR: ……

SIGNATURE OF INVESTIGATOR: …… **DATE:** ...……

Please note: all parties signing the consent form must date their own signature.

Appendix 5 Patient information sheet and consent form for early breast cancer trial

INFORMED CONSENT FORM SAMPLE 1

IBCSG Trial 15-95

Randomized Trial of 3 Cycles of High-Dose Epirubicin + Cyclophosphamide versus 4 Cycles of Epirubicin/Adriamycin + Cyclophosphamide and 3 Cycles of Cyclophosphamide + Methotrexate + Fluorouracil as Adjuvant Treatment for High Risk Operable Stage II and Stage III Breast Cancer in Premenopausal and Young Postmenopausal Patients

Plain Language Statement

Introduction and Aims of the Study

As has already been explained to you, you have a cancer of your breast. This has been treated by surgery and the tissue removed from under your arm has shown that there is evidence of spread of the cancer to the lymph nodes under your arm. Unfortunately, the amount of tissue involved with tumour in your case implies a high risk of the tumor returning.

Recent studies have shown that up to 90 per cent of women (a high risk group) with this problem have the cancer come back either in the breast or elsewhere. This is because some of the cancer cells have escaped into the body before the breast has been removed.

Because of this risk, we are conducting a clinical trial to determine whether the addition of further treatment after surgery decreases the chance of the cancer recurring either in the breast or in the body generally.

We know from previous clinical trials that the use of chemotherapy at standard doses after breast surgery reduces the chance for cancer recurrence and improves survival in patients at lower risk than yourself. In an effort to improve your chances of survival. we are assessing whether giving you higher doses of chemotherapy is a more effective in killing all remaining cancer cells.

In this study, approximately 110 women with breast cancer will receive very high doses of chemotherapy and will be compared with the same number of women with breast cancer who receive standard doses of treatment. This is a randomized study, that is, your treatment will be decided by chance and you have a 50:50 chance of receiving either treatment. You and your treating doctor cannot choose your treatment.

The Design of the Study

HIGH-DOSE CHEMOTHERAPY

If you are assigned to this treatment, you will receive 3 treatments with very high doses of the chemotherapy drugs epirubicin and cyclophosphamide. A problem with giving these doses of chemotherapy is that it lowers your resistance to infection. This lowering of resistance is due to the toxic effects of the chemotherapy drugs on normal cells, particularly the in bone marrow where the blood is made. This can lead to severe infections and prolonged hospital stays for treatment. When the anticancer drugs epirubicin and cyclophosphamide are used at high doses, they must be followed by “rescue” with previously collected blood cells. These cells improve the production of blood which is necessary because of damage to bone marrow cells. In order to collect sufficient numbers of cells from your blood, we give you a hormone called granulocyte colony stimulating factor (G-CSF) that controls the release of the bone marrow or “stem” cells into the blood. Without

the use of G-CSF, there are too few of these “stem” cells in the blood to allow collection of sufficient cells for high dose chemotherapy.

To collect the stem cells from the blood, you will need a catheter (tube) to be inserted via your chest into the big vein that drains blood into your heart. This may be done under a local or general anesthetic. This catheter is used for collecting the stem cells, taking blood tests, and giving chemotherapy and antibiotics or blood or platelets (if needed). The catheter is left in place for the duration of the treatment, about 4 months. Complications can occasionally occur with insertion of the catheter, such as a punctured lung. The catheter may also become blocked or infected, and may need replacing before your treatment is finished.

Prior to receiving chemotherapy, you will undergo collection of blood “stem” cells by a procedure called leukapheresis, performed using a cell separator machine. This equipment is the same as that used by the Red Cross and hospitals to collect plasma or other blood products from patients with normal bone marrow. This will require you to have a needle inserted into a vein in your arm. This procedure takes two to four hours to perform. You will be given G-CSF for 6 days and have the cell collection procedure performed on the 5th, 6th and 7th days. The hormone is given by continuous injection under the skin using a portable pump, which is worn on a belt. You will need to wear this pump for the 6 days. The hormone is given to you as an outpatient, and the cell collection procedures are performed in a day ward. You will only require one period of stem cell collection.

You will then be given 3 treatments with high doses of epirubicin and cyclophosphamide. The drugs are given by injection into the catheter in your chest. Each will be given with the a portion of the blood stem cells that were collected, and further G-CSF treatment. The treatments will be given at 21 day intervals. You will be hospitalized for about 5 days to receive the chemotherapy, and for a variable period after that, depending if you develop an infection or severe complications from the treatment.

Your doctors expect the G-CSF to stimulate the normal bone marrow cells and reduce the risk from infections. No other licensed alternative therapies are known to have this stimulating effect on bone marrow and blood cells. The treatment with the G-CSF allow us to use repeated courses of high dose chemotherapy and so have a better effect on your cancer.

<u>Possible Side Effects</u>

This approach has been tested in over 60 patients in a pilot study in Australia. The treatment was generally well tolerated, and all women recovered from the side-effects of therapy.

Apart from the effects of the chemotherapy drugs on the white blood cells, the other problems encountered in the pilot study included:

a) damage to the red blood cells and platelets, causing anaemia and possible increased risk of bleeding. - you may need a transfusion if this occurs.

b) irritation of the bladder - this risk is reduced by giving you a specific antidote (a drug called MESNA) intravenous fluids before and after chemotherapy.

c) nausea and vomiting - you will be given anti-sickness medications to reduce this.

d) temporary hair loss.

e) soreness in the mouth and throat, including mouth ulcers

f) at very high doses of epirubicin impaired function of the heart can occur, but this is preventable with routine monitoring of your heart function.

g) if you are still having your periods, there is a strong possibility that these will cease after treatment (ie you will become menopausal).

You will be carefully monitored for any side effects during the course of treatment. After receiving the chemotherapy, you will have a blood test daily until after the blood has recovered.

There have been few side effects identified with the use of G-CSF, and the most frequent is mild to moderate aching the bones, which can be controlled with paracetamol, has been reported by about 10 to 20% of patients.

If you or your doctor believe that any of the side-effects from G-CSF and/or chemotherapy become unacceptable you will be withdrawn from the study and treated for your cancer with conventional medications.

STANDARD DOSE THERAPY

If you are assigned the standard-dose treatment, you will receive 1 chemotherapy (drugs called adriamycin and cyclophosphamide, which are injected into a vein once every 3 weeks for four times, for a total of 12 weeks of treatment), followed immediately by the second chemotherapy (drugs called cyclophosphamide, methotrexate and fluorouracil, which are given as tablets daily for 14 days and by injection into the vein on 2 days, every 4 weeks for 3 times, for a further 12 weeks of treatment). This treatment is associated with a similar spectrum of side-effects as the high-dose therapy, however they are likely to be much less severe. All treatment will be administered to you as an outpatient, unless you need to be admitted for a complication. You will be carefully monitored for any side effects during the course of treatment.

BOTH TREATMENTS

Prior to commencing treatment, you will need tests to assess the exact extent of your cancer as would normally be done prior to the start of any chemotherapy - these may include: xrays, CT scans, a bone scan and a heart scan. You will also have blood tests, urine tests, a tracing of your heart (ECG) and a full physical examination to give us a good understanding of your state of health before treatment commences.

There have been 4 cases of cancer of the uterus reported with long term (2 years or more) use of tamoxifen. These women, however, received tamoxifen in double the doses used in this program. Moreover, 200,000 women have been given tamoxifen for more than 2 years in the standard doses used in this study without an apparent increase in the rate of uterine cancer.

The chemotherapy drugs given in this study are associated with a risk of the later development of leukemia. The risk is thought to be < 0.2% when the drugs are given at the standard doses. A recent report form the USA suggests there might be an increased risk of leukemia in patients receiving increased doses of the drugs. In a study with 2548 patients with breast cancer treated with higher than normal doses of cyclophosphamide and normal doses of doxorubicin (a very similar drug to epirubicin) with G-CSF, 5 women developed acute leukemia with features indicating it was caused by the chemotherapy drugs. The risk of leukemia appears to be minor in patients receiving standard doses of chemotherapy, and any increase in this risk with higher doses must be balanced against the possible benefit in terms of improved survival.

In an effort to monitor your ability to cope with both the disease and the therapy, we will ask you to fill out a brief "quality of life" questionnaire at various times. This takes 5 to 10 minutes to complete.

In the first 2 years after therapy you will be seen every 3 months, then every 6 months for 3 years (ie to 5 years), and every year thereafter.

Authorization to Allow Inspection of Medical Reports

It is important for representatives of the health authorities (the National Department of Health) and the hospital ethics committee to be able to inspect your medical records. Any such review will be performed in such a way so that your identity is not revealed. Therefore, you are requested to authorise your doctor to allow such representatives access to review your medical record.

You should ask for any information you want

If you would like more information about the study or if there is any matter about it that concerns you, either now or in the future, do not hesitate to ask one of the researchers or one of the doctors treating you. People you can ask include _____.

Before deciding whether or not to take part you may wish to discuss the matter with a relative or friend or with your local doctor. You should feel free to do this.

Deciding Not to Participate or Withdrawing From the Study

Your participation in this study must be voluntary

It is important that you understand that your participation in this study must be voluntary. This is the case with all research projects in the hospital.

If you do not wish to take part, you are under no obligation to do so. Also, if you decide not to take part, or to withdraw, it will not affect your routine medical treatment or your relationship with those treating you, or your relationship with the Hospital.

Consent Form

I have read the above information and I understand the purpose, benefits and risks of this clinical study and voluntarily agree to participate. My signature below also acknowledges that I have been given a copy of this form for my personal records.

_____ _____
Patient’s Signature Date

_____ _____
Doctor's Signature Date

_____ _____
Witness' Signature Date