Evidence-Based Interpretation Guidelines for Quality of Life Measures

by

Kim Cocks

Submitted in accordance with the requirements for the degree of

PhD

University of Leeds

School of Medicine

July 2011

Intellectual Property and Publication Statements

The candidate confirms that the work submitted is her own, except work which has formed part of jointly-authored publications. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Chapter 3 is published in the European Journal of Cancer; K Cocks, MT King, G Velikova, PM Fayers and JM Brown (2008). Quality, interpretation and presentation of European Organisation for Research and Treatment of Cancer quality of life questionnaire core 30 data in randomised controlled trials. *European Journal of Cancer, Volume 44, Number 13, p. 1793-1798.*

The candidate is first author on the publication and was responsible for drafting the manuscript and submission. The candidate had the original idea to conduct a review of the quality of randomised controlled trials reporting the QLQ-C30 data and conducted all aspects of the systematic review. Prof Brown was the second reviewer for the systematic review. All authors helped develop the idea and reviewed drafts of the manuscript.

Parts of Chapter 7, 9 and the abstract are based on the following jointly authored publication in the Journal of Clinical Oncology; K Cocks, MT King, G Velikova, M Martyn-St-James, PM Fayers and JM Brown. "Evidence-Based Guidelines for the Determination of Sample Size and Interpretation of the European Organisation for the Research and Treatment of Cancer Quality of Life Questionnaire Core 30" *Journal of Clinical Oncology (Jan 2011), Number 29, Issue 1, p. 89-96.*

The candidate is first author on the publication and was responsible for drafting the manuscript and submission. This publication contains the results from the main part of the thesis therefore the candidate led the project design and protocol writing, oversaw the day to day running of the project, extracted and imputed data for the meta-analysis, developed the statistical methods and carried out the statistical analysis. Ms Martyn St-James helped with obtaining the source papers, co-ordinating the expert reviews and entering data onto the database. Prof King had the original idea for evidence-based interpretation guidelines. All other authors helped improve the methodology and were involved in the review of results and interpretation of the analysis.

Abstract

Aim: To use published literature to obtain estimates of large, medium and small differences in quality of life (QOL) data for the European Organisation for the Research and Treatment of Cancer Quality of Life Questionnaire Core 30 (EORTC QLQ-C30).

Methods: An innovative method combining systematic review of published studies, expert opinions and meta-analysis was used to obtain estimates of large, medium and small differences for QLQ-C30 scores. Published mean data were identified from the literature. Differences between groups of patients and over time within patients were reviewed by 34 experts in QOL measurement and cancer treatment. The experts, blinded to QOL results, were asked to predict these differences. Differences were combined using meta-analytic techniques to obtain estimates of small, medium and large effects. Qualitative interviews with patients and experts were used to assess the new methodology.

Results: 911 articles were identified, with 211 relevant articles (3444 contrasts) for the analysis. Our systematic review of the randomised controlled trials (RCTs) showed that the clinical relevance of QOL differences was rarely discussed. Our meta-analysis estimates varied depending on the subscale and on whether QOL was improving or deteriorating. Thus, the recommended minimum to detect medium differences between groups ranges from 7 (diarrhoea) to 19 points (role functioning). When interpreting differences over time a minimum of 7 points represents a medium difference but for most subscales a larger difference is required for a medium deterioration compared with a medium improvement.

Conclusion: Guidelines for interpreting the size of effects are provided for the QLQ-C30 subscales. These guidelines can be used for sample size calculations for clinical trials and to interpret differences in QLQ-C30 scores. The novel methodology was shown to be robust in sensitivity analyses but benefitted from a thorough quality assessment and using only the best quality evidence to derive the guidelines.

Acknowledgements

This research has been carried out by a team which has included myself, Profs Julia Brown (JB), Peter Fayers (PF), Madeleine King (MK), Galina Velikova (GV) and Dr Marrissa Martyn St-James (MMSJ). My own contributions, fully and explicitly indicated in the thesis, have been the project design and protocol writing, day to day running of the project, research assistant staff management, database design, expert review manual and coversheet design, data collection, data extraction/imputation, statistical methodology, statistical analysis and interpretation. I was the Principal Investigator for the patient sub-study which involved writing the protocol and interview format, submission for ethical approval, recruiting patients, conducting the interviews and carrying out the qualitative analysis.

The other members of the group and their contributions have been as follows: JB, MK and PF were joint supervisors and have overseen all aspects of the project including discussion of key issues, advice on statistics and review of results/interpretation. MK also had the original idea for evidence-based guidelines and developed the initial methodology. GV advised on the patient sub-study, conducted expert reviews, reviewed results and helped with interpretation. MMSJ was a research assistant helping with the systematic review and expert reviews.

This project was funded by Cancer Research UK, to whom I am extremely grateful. I also thank a wider team of people who helped make this project a success; Cris Bunker/Will Crocombe (database), Paulina Sieczko (research assistant), Prof Jane Blazeby/Prof Jane Maher (patient interview advice), Dr Penny Wright, Emma Podmore and Lyndsay Campbell who helped with identification of potentially eligible patients and gave advice on gaining informed consent and interview techniques, Dr Pete Wall who kindly volunteered to review the thesis and my Mum for proof-reading. Thanks also to the patients who gave their precious time to participate in the pilot study.

I would like to extend my gratitude to my supervisors and Galina, who to all intents and purposes took on the role of another supervisor. I was extremely lucky to work with such a great team of experts and thank them for their effort and dedication to my research. Without each and every one of them this thesis would not have been possible.

And lastly, but certainly not least, I thank my beautiful family who have shared me with my laptop for too many years now. Thank you to Matthew and Georgie for being such angel babies, and thank you to Dave for doing my share of the housework (most of the time anyway).

Contents

List of Figures

List of Tables

List of Appendices

List of Abbreviations

Acronym	Definition
AP	Appetite Loss
BMT	Bone Marrow Transplant
CF	Cognitive Functioning
CI	Confidence Interval
CO	Constipation
CONSORT	Consolidated Standard of Reported Trials
COPD	Chronic Obstructive Pulmonary Disease
COREQ	Consolidated criteria for reporting qualitative research
CRUK	Cancer Research UK
DI	Diarrhoea
DY	Dyspnoea
EBES	Evidence-Based Effect Sizes
EBIG	Evidence-Based Interpretation Guidelines
ECOG	Eastern Cooperative Oncology Group
EF	Emotional Functioning
EORTC	European Organization For Research And Treatment Of Cancer
ES	Effect Size
FA	Fatigue
FACT-G	Functional Assessment of Cancer Therapy - General
FI	Financial Impact
FU	Follow-Up
GI	Gastro-Intestinal
H&N	Head And Neck
HRQOL	Health-Related Quality of Life
ICC	Intra-Cluster Correlation
ID	Identification Number
KPS	Karnofsky Performance Status Scale
MCID	Minimally Clinically Important Difference
MD	Mean Difference
MeSH	Medical Subject Headings
MIC	Minimally Important Change

MID	Minimally Important Difference
NB	Nota bene
NV	Nausea & Vomiting
PA	Pain
PF	Physical Functioning
PPM	Patient Pathways Manager
PRO	Patient-Reported Outcome
QL	Global quality of life scale
QLQ-C30	Quality of life questionnaire core 30
QOL	Quality of Life
RCT	Randomised Controlled Trial
REML	Residual Restricted Maximum Likelihood
RF	Role Functioning
ROC	Receiving Operator Characteristics
SAS®	Statistical Analysis Software
SD	Standard Deviation
SE	Standard Error
SEM	Standard Error of Measurement
SF	Social Functioning
SL	Sleep
SSQ	Subjective Significance Questionnaire

1 Introduction

1.1 Structure and content of the thesis

This study aimed to produce guidelines for the interpretation of a questionnaire used to measure quality of life in cancer patients, the European Organization for Research and Treatment of Cancer (EORTC) QLQ-C30. The study is referred to as the 'Evidence-Based Interpretation Guidelines' (EBIG) project. Chapter 1 introduces the concept of health-related quality of life and describes the two most commonly used questionnaires in cancer research, the EORTC QLQ-C30 and the Functional Assessment of Cancer Therapy General (FACT-G) questionnaire. Chapter 2 reviews the techniques currently available for interpreting quality of life scores and includes a discussion of methodology developed by King *et al*(1;2) which used published data from the FACT-G to create guidelines for the interpretation of scores 'evidence-based effect sizes' (EBES). Chapter 3 provides the background to the research and highlights the need for interpretation guidelines. This chapter reports a systematic review I carried out to assess the quality of the reporting from randomised controlled trials (RCT) using the QLQ-C30. I also investigated how widely the currently available guidelines for interpreting QOL scores were being used in these papers. This chapter has been published in the European Journal of Cancer(3). The review highlighted the fact that only 38% of papers addressed the clinical significance of the QOL differences and there was a lack of consistency across the papers in how the QOL scores were interpreted.

The main objective of the research was to build on the work of King *et al* and produce interpretation guidelines for the QLQ-C30 questionnaire. A literature search was carried out to identify all articles containing mean scores from the QLQ-C30 questionnaire for a group of patients over time or between groups of patients. The identified papers then underwent a review by a panel of experts in order to group the QOL differences into large, medium, small or trivial categories for a meta-analysis. Chapter 4 details the full methods used to produce the interpretation guidelines and highlights improvements to the original methodology. Chapter 5 summarises the results of the literature search.

Chapter 6 investigates the quality of the meta-analysis. The expert opinion was key to the development of the guidelines and since the methodology was new various

methods were used to ensure their quality, including qualitative and quantitative methods. Chapter 6 is divided into two parts. Part 1 reports on a qualitative study carried out with the expert review panellists and Part 2 uses a quantitative approach to assess the quality of the meta-analysis.

Chapter 7 contains the main study results and interpretation guidelines. We have published the guidelines for comparisons between groups of patients in the Journal of Clinical Oncology(4).

Alongside the main study I carried out interviews with patients aimed at finding out how patients could contribute to and improve the methodology. The results from this qualitative study are reported in Chapter 8.

Conclusions and a discussion of the overall findings can be found in Chapter 9.

1.2 Quality of life (QOL)

1.2.1 QOL definition and instruments to measure QOL

The EORTC describes quality of life assessment in cancer clinical trials as providing "a more accurate evaluation of the well-being of individuals or groups of patients and of the benefits and side-effects that may result from medical intervention."(5) Various definitions of QOL have been proposed and no one concise definition has emerged as a standard. However, most definitions include one or more of the following concepts; health status, physical functioning, symptoms, psychological adjustment, well-being, and life satisfaction(6). Quality of life is defined by Ferrans(7) as "a person's sense of well-being that stems from satisfaction or dissatisfaction with the areas of life that are important to him/her".

Patients' perception of their health and well-being, is an important aspect of health care. Quality of life can be measured using questionnaires which ask a series of questions regarding how the patient is functioning in various aspects of their life or any symptoms they are experiencing. Each question has a selection of possible answers to choose from, e.g. 'Not at all', 'A little', 'Sometimes', 'Very much'. These questionnaires are often used within clinical trials in order to gain experience of the effect of treatments on quality as well as quantity of life. This has become a vital part of cancer research in particular as treatments are often toxic, and the benefits, whether extended survival or palliation of symptoms, are often offset by adverse side effects.

The answers to questions in the QOL questionnaire are used to generate a QOL score for that patient for the time period covered by the questionnaire. These scores

are designed to measure the patients' health and well-being. During 20 years of QOL research, a rigorous methodology has been developed to create the questions and a scoring system so the questionnaires measure QOL reliably and validly. However, like other clinical measurements, the meaning of scores is not clear until the questionnaires have been widely used in different patients in different circumstances. Familiarity with the questionnaire and its range of scores is required to start to understand the clinical significance of QOL scores.

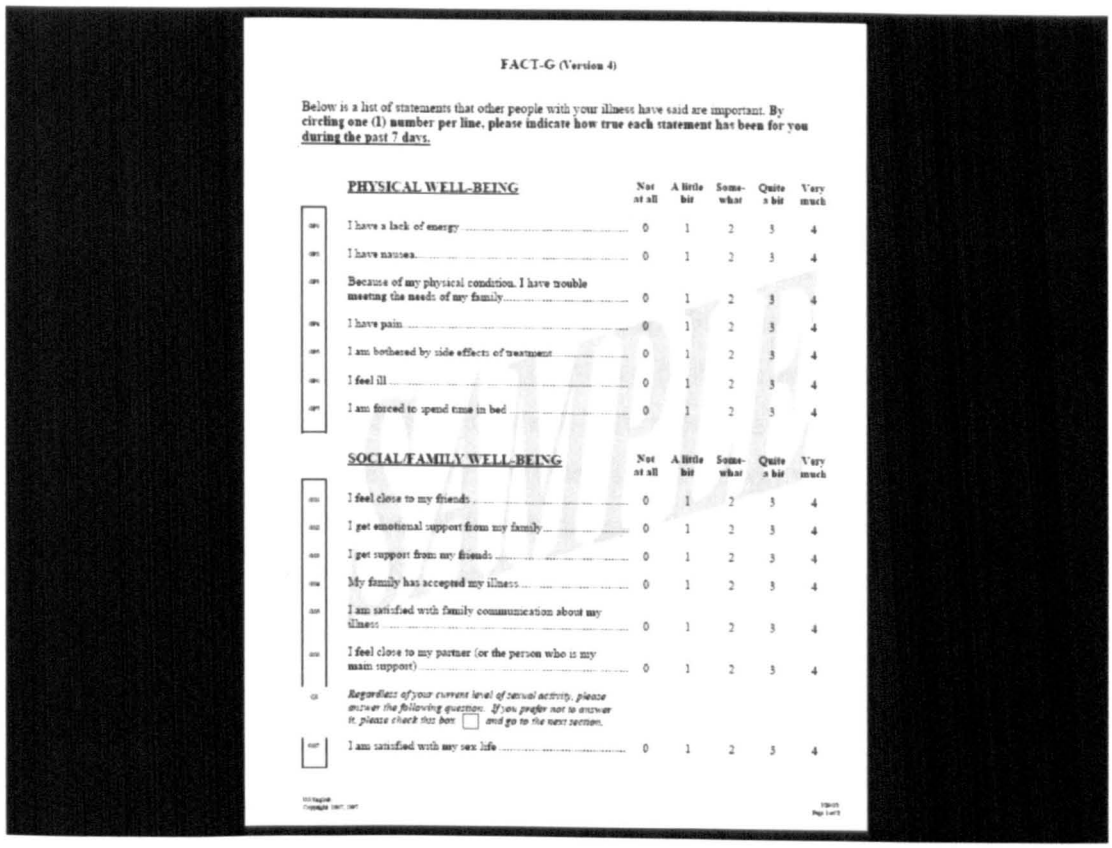
1.2.2 QOL measurement in Cancer

Cancer and its treatment can affect many different aspects of QOL. Initially just the diagnosis can affect a person's life with the associated fear and anxiety it can cause. Then the anti-cancer treatments may have side effects such as nausea, fatigue, vomiting and hair loss. The disease and its treatment can also affect a person's ability to work and socialise as they normally would.

There are two instruments which are commonly used to measure QOL in cancer patients; the EORTC QLQ-C30 and the FACT-G. These account for the majority of published trials of QOL in cancer and are the focus for the thesis. The FACT-G was developed in the USA(8) and the QLQ-C30 questionnaire(9) in Europe.

FACT-G was released in 1993 after 5 years of development and testing(8). It was originally developed as a cancer-specific measure, but its scope was then broadened to chronic conditions generally. The FACT-G (see Figure 1) comprises 27 questions that assess four primary dimensions of QOL: physical (7 items), social and family (7 items), emotional (6 items), and functional well-being (7 items). It uses 5-point Likert-type response categories ranging from 0 = 'not at all' to 4 = 'very much'. The total FACT-G score is the summation of the four subscale scores and ranges from 0 to 108.

Figure 1 FACT-G quality of life questionnaire

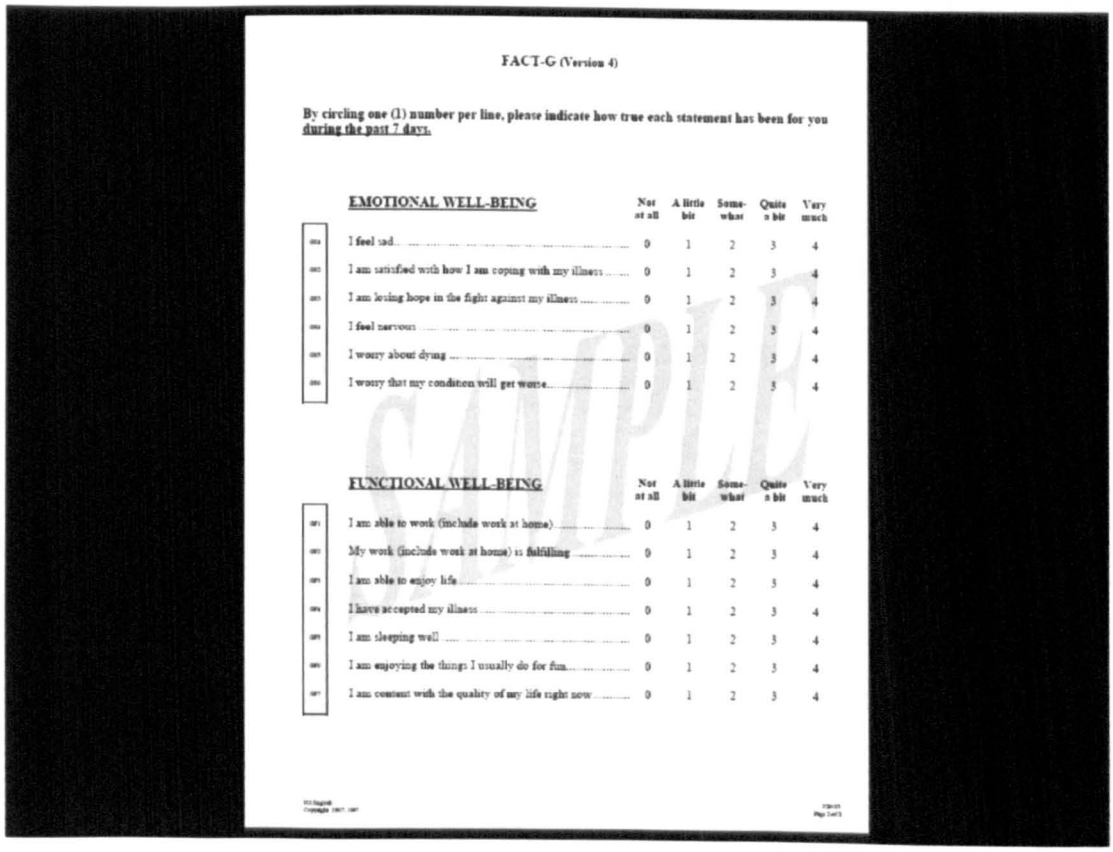


FACT-G (Version 4)

Below is a list of statements that other people with your illness have said are important. By **circling one (1) number per line, please indicate how true each statement has been for you during the past 7 days.**

	PHYSICAL WELL-BEING	Not at all	A little bit	Some-what	Quite a bit	Very much
GP1	I have a lack of energy	0	1	2	3	4
GP2	I have nausea	0	1	2	3	4
GP3	Because of my physical condition, I have trouble meeting the needs of my family	0	1	2	3	4
GP4	I have pain	0	1	2	3	4
GP5	I am bothered by side effects of treatment	0	1	2	3	4
GP6	I feel ill	0	1	2	3	4
GP7	I am forced to spend time in bed	0	1	2	3	4

	SOCIAL/FAMILY WELL-BEING	Not at all	A little bit	Some-what	Quite a bit	Very much
GS1	I feel close to my friends	0	1	2	3	4
GS2	I get emotional support from my family	0	1	2	3	4
GS3	I get support from my friends	0	1	2	3	4
GS4	My family has accepted my illness	0	1	2	3	4
GS5	I am satisfied with family communication about my illness	0	1	2	3	4
GS6	I feel close to my partner (or the person who is my main support)	0	1	2	3	4
GS	*Regardless of your current level of sexual activity, please answer the following question. If you prefer not to answer it, please check this box* ☐ *and go to the next section.*					
GS7	I am satisfied with my sex life	0	1	2	3	4



FACT-G (Version 4)

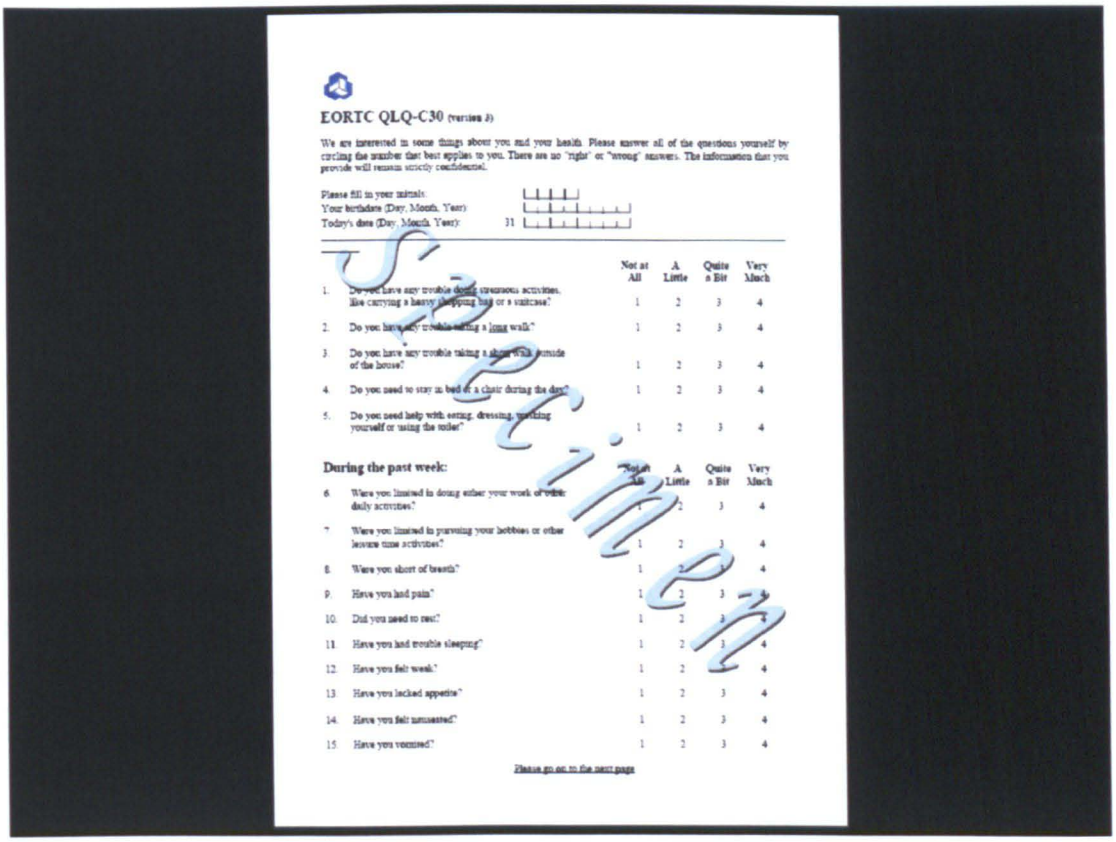
By circling one (1) number per line, please indicate how true each statement has been for you during the past 7 days.

	EMOTIONAL WELL-BEING	Not at all	A little bit	Some-what	Quite a bit	Very much
GE1	I feel sad	0	1	2	3	4
GE2	I am satisfied with how I am coping with my illness	0	1	2	3	4
GE3	I am losing hope in the fight against my illness	0	1	2	3	4
GE4	I feel nervous	0	1	2	3	4
GE5	I worry about dying	0	1	2	3	4
GE6	I worry that my condition will get worse	0	1	2	3	4

	FUNCTIONAL WELL-BEING	Not at all	A little bit	Some-what	Quite a bit	Very much
GF1	I am able to work (include work at home)	0	1	2	3	4
GF2	My work (include work at home) is fulfilling	0	1	2	3	4
GF3	I am able to enjoy life	0	1	2	3	4
GF4	I have accepted my illness	0	1	2	3	4
GF5	I am sleeping well	0	1	2	3	4
GF6	I am enjoying the things I usually do for fun	0	1	2	3	4
GF7	I am content with the quality of my life right now	0	1	2	3	4

The QLQ-C30 consists of 30 questions (Figure 2). The answers to these questions are then scored into 15 subscales measuring different aspects of QOL. The questionnaire is composed of both multi-item scales (i.e. more than one question contributes to an overall score for that scale) and single-item scales. There are five functional scales (cognitive (CF), physical (PF), emotional (EF), social (SF) and role functioning (RF)), three symptom scales (pain (PA), fatigue (FA), nausea and vomiting (NV)), a global health status / QOL scale (QL), and six single items (appetite loss (AP), constipation (CO), diarrhoea (DI), dyspnoea (DY), sleep (SL) and financial impact (FI)). Each of the multi-item scales includes a different set of items and no item occurs in more than one scale. All of the scales and single-item measures range in score from 0 to 100. A high score represents a higher response level. A high score for a functional scale represents a high/healthy level of functioning, a high score for the global health status/QOL represents a high QOL, but a high score for a symptom scale/item represents a high level of symptomatology/problems.

Figure 2 EORTC QLQ-C30 quality of life questionnaire



EORTC QLQ-C30 (version 3)

We are interested in some things about you and your health. Please answer all of the questions yourself by circling the number that best applies to you. There are no "right" or "wrong" answers. The information that you provide will remain strictly confidential.

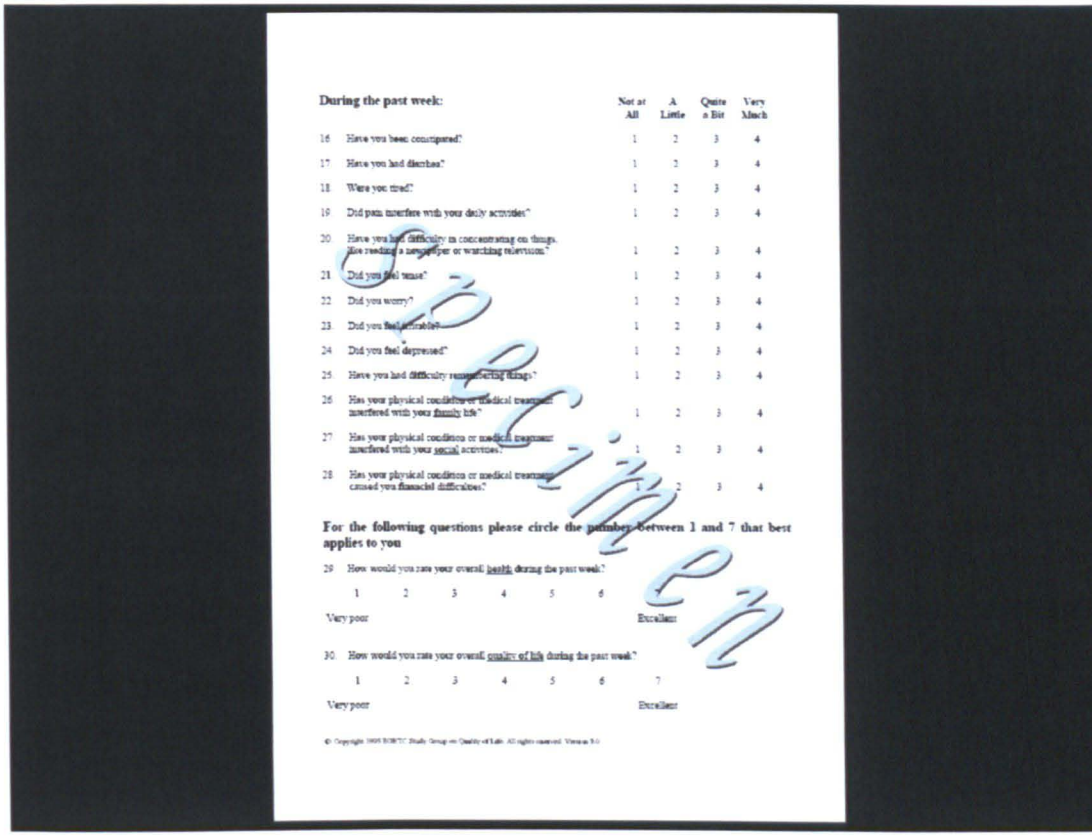
Please fill in your initials:
Your birthdate (Day, Month, Year):
Today's date (Day, Month, Year): 31

		Not at All	A Little	Quite a Bit	Very Much
1.	Do you have any trouble doing strenuous activities, like carrying a heavy shopping bag or a suitcase?	1	2	3	4
2.	Do you have any trouble taking a long walk?	1	2	3	4
3.	Do you have any trouble taking a short walk outside of the house?	1	2	3	4
4.	Do you need to stay in bed or a chair during the day?	1	2	3	4
5.	Do you need help with eating, dressing, washing yourself or using the toilet?	1	2	3	4

During the past week:

		Not at All	A Little	Quite a Bit	Very Much
6.	Were you limited in doing either your work or other daily activities?	1	2	3	4
7.	Were you limited in pursuing your hobbies or other leisure time activities?	1	2	3	4
8.	Were you short of breath?	1	2	3	4
9.	Have you had pain?	1	2	3	4
10.	Did you need to rest?	1	2	3	4
11.	Have you had trouble sleeping?	1	2	3	4
12.	Have you felt weak?	1	2	3	4
13.	Have you lacked appetite?	1	2	3	4
14.	Have you felt nauseated?	1	2	3	4
15.	Have you vomited?	1	2	3	4

Please go on to the next page



During the past week:

		Not at All	A Little	Quite a Bit	Very Much
16	Have you been constipated?	1	2	3	4
17	Have you had diarrhea?	1	2	3	4
18	Were you tired?	1	2	3	4
19	Did pain interfere with your daily activities?	1	2	3	4
20	Have you had difficulty in concentrating on things, like reading a newspaper or watching television?	1	2	3	4
21	Did you feel tense?	1	2	3	4
22	Did you worry?	1	2	3	4
23	Did you feel irritable?	1	2	3	4
24	Did you feel depressed?	1	2	3	4
25	Have you had difficulty remembering things?	1	2	3	4
26	Has your physical condition or medical treatment interfered with your family life?	1	2	3	4
27	Has your physical condition or medical treatment interfered with your social activities?	1	2	3	4
28	Has your physical condition or medical treatment caused you financial difficulties?	1	2	3	4

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

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

1 2 3 4 5 6 7

Very poor — Excellent

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

1 2 3 4 5 6 7

Very poor — Excellent

1.3 Interpretation of QOL scores

Although QOL questionnaires are now widely used in cancer research, their potential to impact on treatment practice cannot be fully realised until the scores and changes in these scores can be interpreted. Interpreting the clinical significance of effects observed on QOL scales is problematic because their units of measurement are unfamiliar to clinicians, policy makers and patients alike (10). For example, say a clinical trial found an average difference of 20 points between two treatment groups, indicating that Treatment A had better quality of life than Treatment B. However, Treatment B was found to improve survival by a small amount. In order to inform clinical practice, clinicians need to know if this size of change in QOL is meaningful to patients or worthy of clinical attention. Until the QOL scores can be interpreted it is not possible to weigh up the survival advantage versus the QOL disadvantage and make informed decisions. There is a pressing need to make the results of QOL assessments more clinically interpretable so they can be more informative in practice.

QOL instruments undergo extensive testing to ensure their validity and responsiveness. For example, known-group comparisons(11) test whether the questionnaire can distinguish between two groups known to be clinically different. Analyses to look at how responsive the questionnaire is over time may also be

conducted. As part of this validation process some limited data on interpretation of scores emerges but the process does not address what the smallest change in score is that would be meaningful to patients or clinicians. Meaningful changes may be those that lead to a change in a patient's daily life or that lead to a change in patient management and, importantly, the degree of change deemed 'meaningful' may differ with perspectives. Various methods for trying to interpret QOL scores have been explored and these are reviewed fully in Chapter 2. Several authors provide a comprehensive overview of existing interpretation strategies(12-15).

1.3.1 Minimally important differences

The interpretation of changes (or differences in scores) is frequently based on the minimally important difference (MID). Although the precise definitions of MID vary it is essentially the threshold that separates trivial differences from those that, although small, are important. King(16) cites the definition from Jaeschke *et al*(17) as probably the most influential in the field. They defined the minimally clinically important difference (MCID) as "the smallest difference ... which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management".

De Vet *et al*(18) highlight the importance of also distinguishing between the minimally detectable difference (i.e. the smallest change the instrument can detect) and the minimally important difference as described above. Both are relevant to the interpretation of scores, particularly for individual patients (since it is important to know the MID and also if the instrument is capable of detecting a change as small as the MID).

1.3.2 Methods of interpretation

The approaches to interpretation are broadly categorised into distribution-based or anchor-based(10). Distribution-based methods use statistical parameters to establish meaningful differences based only on observed HRQOL results, while anchor-based methods compare, or anchor, QOL differences to other clinical differences whose interpretation is known or to patients' perception of change. Current recommendations(15;18) suggest that interpretation should be based on a combination of methods but that patient-based and clinical anchors are primary, since they address the importance of differences.

King *et al* recently developed a novel method for developing evidence-based interpretation guidelines, and illustrated it using the FACT-G(1). An early version of this

method was also used to develop preliminary interpretations and evidence-based effect sizes for the EORTC QLQ-C30(19;20). This method aims to collect published data on scores from the questionnaire and use these to develop guidelines for interpretation. Since the QLQ-C30 and FACT-G have been in circulation for a number of years and are the most commonly used cancer-specific QOL instruments, they are good candidates for such a method, with an abundance of published data from many different kinds of studies.

The strength of this methodology is that guidelines are developed specifically for a questionnaire using data from that questionnaire. Scores from groups of patients or from a group of patients over time are extracted from the literature. Expert opinion is then used to group them into large, medium, small and trivial differences to gain an estimate of the range of scores that represent a meaningful change in QOL. These estimates would then be published for each scale in the questionnaire as a guide to the difference in QOL scores that may be interpreted as small, medium or large.

Interpretation guidelines of this kind are aimed at comparing groups of patients such as those produced in clinical trials and health services research rather than investigating individual patient changes. Distinguishing between the two types of changes is important since a meaningful change when considering a group of patients is likely to be smaller than for an individual(12). This is because a small change that looks insignificant for a patient may actually translate into an important improvement in the population when considering the same change as an average score instead of an individual score. Although some patients will have an even smaller change than the mean others will have better changes (and some of these may be substantially better than the mean) and this could be a clinically relevant result overall.

The planned guidelines will therefore be of use in planning and interpreting clinical research about the effects of cancer and its treatment on patients' QOL. When planning a clinical trial, investigators require *a priori* knowledge of what constitutes a clinically important effect so that they can calculate the sample size required for the trial. This information is currently unavailable despite substantial experience with cancer QOL instruments. Methods currently used to plan sample size may be under- or over-estimating the required study size, as noted by King *et al(1)*. When a study is complete, both investigators and end-users need to understand the clinical relevance, and hence policy relevance, of the outcomes. Thus the interpretation guidelines and the more general methodology for developing them will facilitate better research about

the QOL of cancer patients, and better understanding of the results and implications of such research.

1.4 Objectives

The main objectives for this thesis were:-

- To further develop and extend the methodology for producing interpretation guidelines using published literature;
- Apply the methodology to create guidelines for interpreting the size of changes in QOL scores from the QLQ-C30.

Chapter 4 describes the methods for this project in full, including the numerous methodological refinements and extensions to the prototype methodology developed by King *et al(1;2)*. In summary: we obtained expert opinion from a much larger panel of experts and targeted their reviews to the cancer types or treatments they specialised in; we used a different scale to collect the expert opinion which allowed for uncertainty in their judgments of the QOL changes; we tried to include as much of the identified literature in the expert review as possible then used statistical analysis to identify the 'best' evidence to go forward into the meta-analysis.

The resulting evidence-based interpretation guidelines can be found in Chapter 7, providing researchers for the first time with separate guidelines for each subscale in the QLQ-C30. Researchers can now more accurately calculate sample size according to the subscale of primary interest and interpret QOL differences using our guidelines, which distinguish between differences between groups of patients and differences observed over time. These guidelines should be more widely applicable than those currently available as they are based on a wide range of cancers and clinical situations.

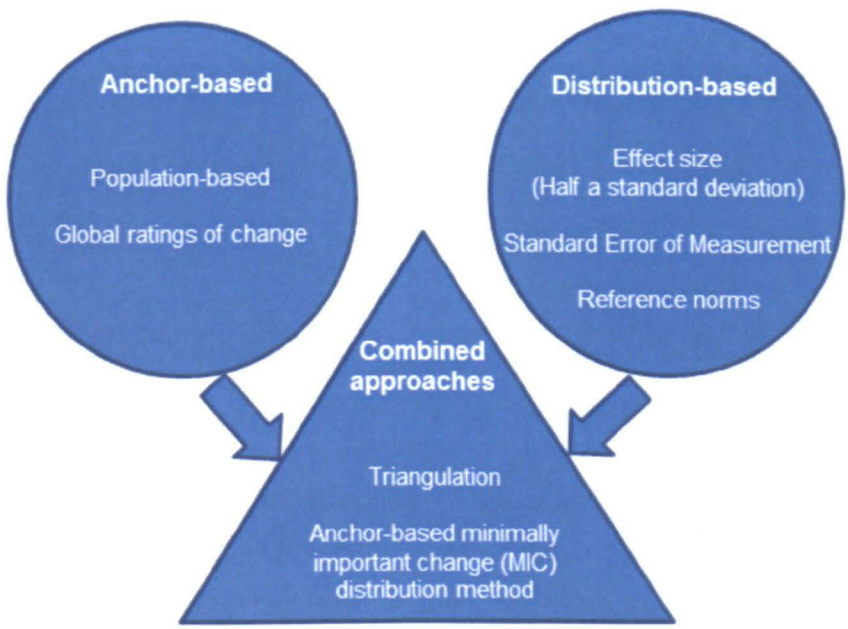
2 Literature review

2.1 Approaches to interpretation

Several authors provide a comprehensive overview of existing interpretation strategies(12-15) and these show a number of different approaches have been used to try to interpret QOL scores by finding the MID(15). However, we have shown through a systematic review of randomised controlled trials reporting data from the QLQ-C30, that no single method has emerged as a standard for interpretation(3).

The approaches to interpretation are broadly categorised into anchor-based or distribution-based methods(10). Figure 3 shows a summary of the methods and they are described in detail in the following sections.

Figure 3 Methods used to interpret QOL scores



2.1.1 Anchor-based methods

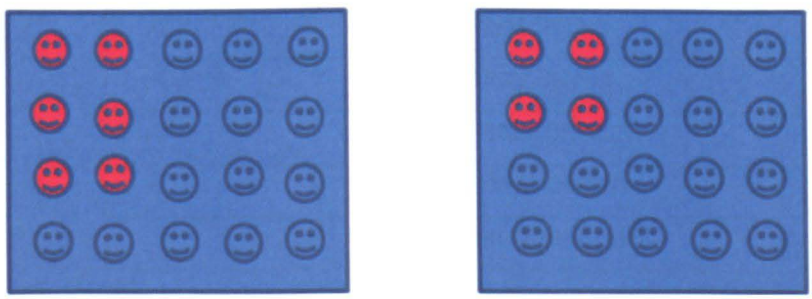
Anchor-based methods relate changes in QOL scores to clinical status. The idea being that an independent measure whose interpretation is well understood is used to anchor QOL scores whose interpretation is as yet unknown. These methods "require

an independent standard or anchor that is itself interpretable and at least moderately correlated with the instrument being explored."(12).

2.1.1.1 Cross-sectional anchors

Wyrwich *et al*(21) describe the use of 'population-based' anchors where differences are expressed in terms of their impact on populations. For example, Group A has a mean QOL score of 50 points compared to Group B with 70 points. The absolute difference of 20 points may be hard to interpret where the QOL scale is largely unknown. Therefore instead of using the 20 points to interpret the difference an anchor (with known meaning) is used. Physical function may be an example of a population-based anchor. If we knew that 30% of Group A patients had quite a bit of difficulty taking a short walk outside of the house compared to 20% of patients in Group B, then there is an absolute difference of 10% or a relative increase of 20% (10/50) with more difficulties in Group A. Figure 4 shows this example in a picture, with red people used to indicate those in the group with quite a bit of difficulty taking a short walk in the two groups.

Figure 4 Population-based anchor example



QOL=50 QOL=70

The authors describe the strengths of these anchors as follows:- "they retain the underlying complexity of the QOL construct that they anchor by providing a probabilistic relationship to a concrete indicator of the underlying concept. Moreover, if multiple population-based anchors are provided for interpreting score changes within a population, individuals can choose the external standard of greatest relevance or base decision on the average of several." So in the above example we could also look at the proportion of patients with the ability to work, proportion with severe pain and so on in order to build up a clearer picture of the meaning of the change in QOL score.

2.1.1.2 Anchors over time

Anchors over time (longitudinal anchors) can also be used rather than anchoring between groups as described above. The global rating of change is the most common anchor used over time to determine the MID(17;22). The global rating of change is measured on a 7 point scale in either direction (Table 1).

Table 1 Global rating of change

Score	Description
-7	A very great deal worse
-6	A great deal worse
-5	A good deal worse
-4	Moderately worse
-3	Somewhat worse
-2	A little worse
-1	Almost the same, hardly any worse at all
0	No change
1	Almost the same, hardly any better at all
2	A little better
3	Somewhat better
4	Moderately better
5	A good deal better
6	A great deal better
7	A very great deal better

Patients complete a QOL instrument over time and use the global rating of change alongside it to assess the degree of change they feel. The score on the global rating of change scale is then used to attach meaning to the QOL scores over the same time period. Scores of 0, 1 or –1 are considered as unchanged, scores of ±2 and ±3 are considered the minimally important difference, scores of ±4 and ±5 are a moderate difference and ±6 and ±7 are considered to be large changes on the QOL scale.

There are however some discrepancies among authors over the size of change on the global rating scale to use as the smallest difference of interest and therefore the size of the MID varies. For example, a change of two to three on the scale in either

direction was used by Juniper *et al*(22) as the MID group compared with a change of one to three used by Jaeschke *et al*(17).

One criticism of these methods is that the global rating of change scale is a single, non-validated item. This is then being used to anchor QOL questionnaires which are usually multi-item and well-validated. Also, because individual patients complete the questionnaires over time, the MID is determined within a specific group of patients therefore it's applicability across disease areas and between groups of patients with different characteristics is questionable.

The reliability of patients' estimates of previous health status is also an issue. The strengths and weaknesses of the global rating of change scale were reviewed by Kamper *et al*(23). They indicate that recall bias(24) and response shift(25) affects whether the ratings actually measure the transition between the two time points as intended. Studies have now found that actually the global rating of change relates more strongly to another patient-reported measure at the same time than to the change in that same measure over the time period in question(26;27).

There is also literature on an adaptation of this method using between-patient ratings of change instead of within-patient ratings (17;28;29). The advantage of using between-patient ratings is that it avoids the issue of recall bias and response shift as patients are asked to compare themselves to others with the same disease rather than remembering changes in their own health over time. The MID results using the between-patient or the within-patient ratings actually resulted in similar MIDs.

2.1.2 Distribution-based methods

Methods in this category use the distribution of QOL scores to interpret meaningful change. The methods generally find the minimally detectable difference which is not the same as the minimally important difference(18). The MID requires input from patients or clinicians. The distribution-based methods result in the change or difference being expressed as a standardised measure rather than trying to estimate the MID. The standardised measure can then be used to estimate the relative size of differences. Some work however has also been done to link the distribution-based estimates with MID or MCID(30).

Guyatt *et al(12)* provide a full description of the available distribution-based methods. The methods which have been used in the interpretation of QOL measures are described in more detail here.

2.1.2.1 Effect size (ES)

One of the earliest distribution-based methods uses effect sizes. The effect size is calculated as:-

$$ES = \frac{Mean\ change\ in\ QOL\ scores}{Standard\ deviation\ at\ baseline}$$

Effect sizes from QOL measures are generally compared with those defined by Cohen(31) as representing small, moderate and large effects. Cohen proposed small, medium and large effect sizes as 0.2, 0.5 and 0.8, respectively. His intention was for these guidelines to better inform sample size calculations. The guidelines are now widely used, not only to calculate sample sizes, but also to estimate a difference or change that might be meaningful. However, Cohen described his guidelines as "arbitrary conventions ... recommended for use only when no better basis for estimating the effect size is available"(31).

2.1.2.2 Half a standard deviation

Others have commented that half a standard deviation seems to have a general universality about it and that Cohen's guidelines for 0.5 as a moderate effect were actually very intuitive. Norman *et al*(32) conducted a review of the literature calculating MID for health-related QOL measures. They studied the magnitude of the MIDs found in 38 studies and concluded that they were all consistently close in size to half a standard deviation. This implies that, regardless of the instrument or disease, half a standard deviation could be considered as a rule of thumb to ascertain important changes. While the authors acknowledge that this should not be viewed as a fixed benchmark they do provide evidence based on psychological theory as to why half a standard deviation may be important.

There are however some criticisms of this work(33;34). The studies compared used different definitions of MID so essentially they were measuring slightly different concepts. There were also some studies excluded from the review because the effect sizes were substantially different to those found in the other studies, therefore the claim that 0.5 standard deviations (SD) was found to be remarkably consistent across the studies is misleading. Although a simple universal rule is attractive for interpretation it seems unlikely that one definition of a meaningful difference exists across all types of diseases and HRQOL aspects.

2.1.2.3 Standard Error of Measurement (SEM)

The SEM is described by Wyrwich (30) as follows, "If a single patient completes the same HRQOL or health status measure repeatedly, with no change in HRQOL or health status taking place between testings and no memory of question and/or response effects, the standard error of measurement reflects the standard deviation of the distribution of his/her repeated questionnaire scores."

The SEM takes into account the reliability (test-retest reliability measured using intra-class correlation coefficient) of the QOL measure as well as within-person variability(SD). It is calculated as follows:

$$SEM = SD\sqrt{1 - reliability}$$

However, there is a lack of agreement on the threshold value for SEM that represents a meaningful difference, with papers showing anything from 1(35) to 2.77(30) as a significant change. There are also discrepancies on how to calculate the reliability of the QOL instrument for use in the calculation of SEM.

Wyrwich *et al*(35) showed that a one-SEM change broadly conformed with Cohen's definition of small, moderate and large effect sizes. The more reliable items had 1-SEM corresponding to Cohen's definition for small effect sizes and the more unreliable items had 1-SEM more similar to Cohen's moderate difference. The authors have also conducted further reviews(30) to try to establish the link between the SEM and the MID but one SEM was not consistently found to represent the MID, with some studies showing up to 2.3 SEMs were equivalent to the MID.

2.1.3 Combined anchor and distribution-based approaches

Anchor-based methods do not take into account the measurement precision of the instrument so cannot account for changes due to random variation alone. The distribution-based methods, however, use only the statistical properties of the sample or the instrument so there is then difficulty in defining what a meaningful change is to patients or clinicians. Therefore more recently attempts have been made to use a combined approach to the interpretation of QOL scores thus using the advantages of both of these approaches.

An early method from Cella *et al*(36) calculated both anchor-based and distribution-based estimates of meaningful change but did not attempt to produce a combined estimate. Jacobson *et al*(37) improved on this method by requiring a change to meet both the criteria from the anchor-based and distribution-based methods before

it could be considered as an important change. Crosby *et al*(38) were the first to try to integrate the methods by describing how to resolve any discrepancies between the estimates from the two methods and also accounting for the fact that MID varies according to the baseline value. Their approach was fairly simple though in that either the anchor-based or distribution-based cut-off was used depending on which was highest. Yost *et al*(39) instead describe using the range of MIDs from different approaches as guidelines.

A different approach to combining any number of anchor and distribution-based methods is through triangulation. Denzin(40) describes methodological triangulation, where multiple methods are used to examine social phenomenon. The rationale being that "the flaws of one method are often the strengths of another: and by combining methods, observers can achieve the best of each while overcoming their unique deficiencies". The MID estimates from global ratings of change, statistical distribution estimates and qualitative data from patients were combined in this way by Leidy and Wyrwich(41). Revicki *et al*(42) describe triangulation as "examining multiple values from different approaches and converging on a small range of values". They suggest plotting the range of MIDs from the different methods on a graph in order to narrow these down to a smaller range or single MID point. They also recommend weighting the approaches though with clinical anchor methods taking precedence.

In a more recent method the integration of the two methods has been more sophisticated. De Vet *et al*(43) developed a visual method called the anchor-based minimally important change (MIC) distribution method. They classified patients into three groups using a clinical anchor; patients with an important improvement, no change or an important deterioration in QOL. The distribution of the change scores for each group were overlaid on a graph. Cut-offs between the size of changes were then decided using either ROC (receiving operator characteristics) optimal cut points or the 95% limit of the distributions. The ROC cut-off is the point that minimises the incorrect classifications (e.g. where patients with no change would be classified as improved or vice versa).

2.1.4 Reference Populations or Norms

QOL data from a study sample can be compared with data from the general population in order to gauge the size of changes. Differences from the general population matched by age and gender or comparison of patients with percentiles from the general population could be used. Differences between reference data sets from different countries however have been found, highlighting the need for data from the

same country as the QOL data. Therefore, in order for these to be a useful interpretation tool data are needed from a large number of countries. The general population has problems in the same domains as those relevant to cancer patients, emphasising that reference population data does not provide a comparison with no symptoms or perfect health but reflects the chronic conditions of aging in a population. Also, this method of interpretation does not address the importance of changes on the QOL scale(44). Section 2.2 highlights the population-based reference values available for the QLQ-C30.

2.2 Existing interpretation strategies for the QLQ-C30

Various methods have been adopted to aid interpretation of the QLQ-C30 specifically. King(20) used 14 studies to estimate effect sizes using clinical anchors and compared these to Cohen's guidelines. They found that the guidelines were approximately adequate for physical, role and symptom scales but not for global and psychosocial dimensions of the QLQ-C30, which resulted in smaller effect sizes.

Population-based reference values have now been published for German(45), Danish(46), Norwegian(46;47) and Swedish(48) populations.

Osoba *et al*(49) published small, moderate and large changes in scores from the QLQ-C30 based on global ratings of change using the Subjective Significance Questionnaire (SSQ). The SSQ used a 7-point scale ranging from much worse through no change to much better. This aimed to address the importance of changes to patients. These differences were 5-10 for a small difference, 10-20 for a moderate difference and >20 for a large difference. However, the SSQ was not found to correlate well with the QLQ-C30 although there was a linear trend between QLQ-C30 scores and the SSQ. This study was carried out in breast cancer and lung cancer patients and it is not clear how the results may generalise for other patient populations. The study was also carried out for specific subscales of the QLQ-C30 (global, physical, emotional and social functioning) and therefore these differences may not be applicable to other subscales. The observations on size of change in score from this study were however similar to those from King *et al*(20).

Each of these methods used to aid interpretation of the QLQ-C30 suffers from shortfalls as highlighted, although they have encouraged clearer presentation of QLQ-C30 results and taken important steps towards starting to interpret rather than simply report QOL results.

3 Background

This chapter is published in the European Journal of Cancer. Kim Cocks, Madeleine T. King, Galina Velikova, Peter M. Fayers and Julia M. Brown (2008). Quality, interpretation and presentation of European Organisation for Research and Treatment of Cancer quality of life questionnaire core 30 data in randomised controlled trials. European Journal of Cancer, Volume 44, Number 13, pages 1793-1798.

The aim of this review was to look at the quality of the reporting of studies using the QLQ-C30 and investigate how widely the available methods for interpreting QOL scores were being used.

3.1 Abstract

Aim: To review reporting standard, presentation and interpretation for quality of life (QOL) outcomes in randomised controlled trials (RCTs) using the EORTC QLQ-C30.

Methods: Cancer RCTs reporting EORTC QLQ-C30 data were identified and reviewed against a reporting quality checklist. Interpretation/presentation methods for QOL data were also recorded.

Results: Eighty-two papers were reviewed. 70% met criteria for high quality reporting; 94% reported mean scores; 84% presented results in tables/graphs; 80% reported p-values or statistical significance. Clinical significance was addressed in 38%. Where clinical significance was not addressed, reliance was usually on statistical significance to interpret the results.

Discussion: EORTC QLQ-C30 results are generally reported well, although it was common to rely on statistical significance alone for interpreting results. While interpretation in terms of clinical significance has improved in recent years, there is still a lack of robust clinical interpretation of QOL results even in papers reported to a high standard.

3.2 Introduction

While a considerable body of evidence about health-related QOL is accruing from cancer clinical trials, the extent of its impact on clinical practice is unclear. One of the

barriers is poor communication of the clinical relevance of the results(50). Reviews of prostate cancer trials(51), breast cancer trials(52), and surgical oncology trials(53), report that 11%, 33% and 67% were able to inform clinical decision-making. Part of this variability in these estimates arose because of differences in how the ability to inform decision-making was measured. Regardless of this, it is clear that in order to inform clinical decision making, a QOL study needs to be designed robustly, reported adequately and interpreted appropriately.

The CONSORT statement(54;55) provides a checklist for reporting RCTs. Efficace *et al.*(56) propose a checklist specifically for evaluating QOL outcomes, listing criteria for reporting QOL outcomes and identifying the essential issues to be addressed in order for a trial to have reliable QOL outcomes. The checklist comprises 11 items grouped into four categories: conceptual, measurement, methodology and interpretation. The authors define a paper with high quality QOL outcomes as one that meets at least 8 out of the 11 criteria and these have to include three high-priority concerns ("baseline compliance reported", "psychometric properties reported" and "missing data documented").

Osoba *et al*(57) and Guyatt/Schunemann(50) recommend that the presentation of QOL data include proportions of patients reporting a QOL benefit. They argue that this provides results meaningful to clinicians and therefore the results are more likely to influence clinical decision-making. Osoba *et al*(57) recommend 10% of the scale as the cut-off point to define improvement, with a stipulation that this degree of change should persist for a reasonable period. As an additional guide to interpretation Guyatt *et al*(50;58) also show how to generate the number needed to treat for one patient to benefit from therapy.

A number of different approaches have been used to develop interpretation for QOL scores. Some are entirely data driven and some use clinical anchors to interpret differences (over time or between groups). However, there are a number of shortfalls of the current methods and no single method has emerged as a standard for interpretation. Some are not specific to the QOL instrument being used and the validity of these is rarely tested for the specific instrument prior to relying on them for interpretation. It is also common to rely on statistical significance in order to interpret whether differences in scores are clinically significant. However, statistical significance does not necessarily imply a meaningful difference in a clinical context, particularly if the minimum clinically important difference in QOL was not determined *a priori* and used to determine the sample size for the trial.

Several authors provide a comprehensive overview of existing interpretation strategies(12-14;59). Various methods have been used to aid interpretation of the EORTC QLQ-C30 specifically, which are of interest for our review. King(20) used 14 studies to estimate effect sizes (mean difference divided by standard deviation) using clinical anchors, and compared these to Cohen's guidelines(31) which propose small, medium and large effect sizes are 0.2, 0.5 and 0.8 respectively. King's estimates were about the same as Cohen's guidelines for physical, role and symptom scales, although for the global and pyschosocial dimensions of the QLQ-C30, the estimates were smaller. Osoba *et al*(49) provided estimates for small, moderate and large changes in scores from the EORTC QLQ-C30 based on retrospective global ratings of change, an approach based on individual patient's rating the importance of changes in QOL. These were found to be 5-10 for a small difference, 10-20 for a moderate difference and >20 for a large difference, similar to those yielded by King's analysis(19;20). This is the basis of Osoba *et al*'s(49) recommendation of 10% of the scale as the cut-off point to define a clinically important change.

This review summarises the quality of QOL reporting for cancer RCTs using the EORTC QLQ-C30 and looks at the methods used to present and interpret the QOL data, particularly in papers that score well on the quality checklist. The interpretation methods used across studies are reviewed to assess how widely used current methods are, the extent to which clinical significance is addressed and whether there is a need for additional interpretation guidelines for the EORTC QLQ-C30. A summary of the methods of presenting the data is used to assess whether it is reported clearly and in a way that may be utilised by clinicians.

3.3 Methods

I searched for potential sources of EORTC QLQ-C30 scores using Cinahl, Medline, Embase, Medline-in-process and Psychinfo concurrently via the Ovid interface. The search terms were qlq c30, quality of life questionnaire c 30, quality of life questionnaire c30, eortc qlq-c30, qlq c33, qlq c30+3. References from the EORTC bibliography(60) were also added. I removed duplicates in Ovid or subsequently in Reference Manager.

I reviewed papers identified as cancer RCTs using the EORTC QLQ-C30 and classified them according to the minimum standard checklist for evaluating QOL outcomes(56). If more than one paper was identified reporting QOL results from the same study then the study report was included rather than a paper reporting any wider

issues, for example comparisons of statistical methods. Papers were defined as having high quality QOL outcomes if they met at least 8 out of the 11 criteria, including the three high-priority concerns ("baseline compliance reported", "psychometric properties reported" and "missing data documented"). If any items were evaluated as "not applicable" then these items were excluded in the evaluation of high quality. A second reviewer (Prof Brown) independently classified a sample of the papers.

Methods used for presenting the data were recorded for all papers in terms of whether text, tables or graphs were used and the type of data presented (e.g. means, medians, effect size, p-values, proportion improved etc.). For papers addressing clinical significance, the method of interpretation was recorded.

3.4 Results

3.4.1 Identification of papers

As of February 2006, 911 papers had been identified using the search strategy. Papers were included if they presented any data from the EORTC QLQ-C30, were cancer trials and were available in English. Ninety-two papers were identified as cancer RCTs reporting EORTC QLQ-C30 scores.

Ten papers were subsequently excluded from the review. Nine used the data to investigate different statistical techniques, conducted extra analyses or reported long-term follow up rather than reporting the results from the RCT and the checklist does not seem applicable. Four of these papers also had a trial report for the same study which was included. One paper was excluded as it was a pilot study with a randomised design exploring the feasibility of QOL assessment in a further RCT. Eighty-two RCTs were therefore included in this review.

3.4.2 Study characteristics

Table 2 shows the characteristics of the 82 studies. Sample sizes ranged from 31 to 1491 patients, with a mean of 272. In total, the papers reported on more than 22000 patients. The majority of studies involved patients with breast cancer (21%), mixed cancer sites (18%), lung cancer (17%) and colorectal cancer (11%). No other cancer sites represented more than 10% of the sample. Patients were from a wide range of countries, although the majority of studies were European (67%).

Table 2 RCT Study characteristics

	Number of studies (%)
Design	
Phase II	7 (9%)
Phase III	25 (30%)
Not specified	50 (61%)
QOL endpoint	
Primary	12 (14%)
Secondary	67 (82%)
Not specified	3 (4%)
Number of patients	
Mean (standard deviation)	272 (239.1)
Median (range)	208 (31-1491)
Region/Country where study conducted	
Europe	
Multi-country	6 (7%)
Austria	1 (1%)
Belgium	1 (1%)
Denmark	2 (2%)
France	4 (5%)
Germany	3 (4%)
Italy	6 (7%)
Netherlands	6 (7%)
Norway	6 (7%)
Spain	1 (1%)
Sweden	8 (10%)
UK	11 (13%)
International	
Multi-country	10 (12%)
Australia	2 (2%)
US/Canada	15 (18%)

	Number of studies (%)
Cancer site	
Breast	17 (21%)
Mixed sites	15 (18%)
Lung	14 (17%)
Colorectal	9 (11%)
Leukaemia/Lymphoma	7 (9%)
Prostate	7 (9%)
Brain	3 (4%)
Oesophageal/Stomach	3 (4%)
Gastro-intestinal	2 (2%)
Malignant melanoma	2 (2%)
Head and neck	1 (1%)
Ovarian	1 (1%)
Testicular	1 (1%)

3.4.3 Quality of reporting

Table 3 shows the level of QOL reporting according to the checklist. As the EORTC QLQ-C30 is a generic instrument designed for all cancer patients and was previously validated in cancer patients, the measurement criteria in the checklist were satisfied for all papers. The main failings of studies were that there was no rationale for using the questionnaire, no details of the administration and not addressing clinical significance of results. The majority of papers (>90%) reported the hypothesis (or stated QOL as an endpoint), stated the timing of assessments and included some general presentation of the results. Fifty-seven (70%) papers met the criteria for high quality. Reporting was of slightly higher quality in the 41 papers whose primary endpoint was QOL or which reported QOL as the main purpose of the paper rather than reporting the overall results of the trial. Thirty-five (85%) of these papers met the criteria for high quality. However, despite QOL results being the main aim of these papers, still only 22 (54%) addressed clinical significance.

Table 3 Level of reporting according to the minimum standard checklist for evaluating QOL outcomes in cancer clinical trials

QOL Issue	All RCTs N=82	High quality N=57	QOL as primary outcome/ aim N=41
Conceptual			
A priori hypothesis stated	77 (96%*)	57 (100%*)	40 (98%)
Rationale for instrument reported	25 (30%)	19 (33%)	18 (44%)
Measurement			
Psychometric properties reported	82 (100%)	57 (100%)	41 (100%)
Cultural validity verified	82 (100%)	57 (100%)	41 (100%)
Adequacy of domains covered	82 (100%)	57 (100%)	41 (100%)
Methodology			
Instrument administration reported	39 (48%)	33 (58%)	29 (71%)
Baseline compliance reported	62 (76%)	57 (100%)	36 (88%)
Timing of assessments documented	82 (100%)	57 (100%)	41 (100%)
Missing data documented	64 (78%)	57 (100%)	38 (93%)
Interpretation			
Clinical significance addressed	31 (38%)	25 (44%)	22 (54%)
Presentation of results in general	78 (95%)	56 (98%)	41 (100%)

*Not applicable for two studies

3.4.4 Presentation of QOL data

Thirty-two (39%) papers used a combination of tables and graphs to summarise the data. Thirteen (16%) used graphical summaries alone and 24 (29%) used tabular displays. Thirteen (16%) reported QOL results in the text with no graphical or tabular summary. A higher proportion of the papers meeting the standard of high quality reporting used both graphs and tables 27 (47%) to display the results.

Fifteen (18%) of papers report the percentage of patients with improved QOL scores as recommended by Osoba *et al*(57) and Guyatt *et al*(58) and this percentage is similar in the subgroup of high quality papers (21%). The definition of 'improvement' varied between the reports however, with some papers using >10 points as an improvement (with or without a minimum length of time for this to be sustained) and

other papers regarding any increase in scores as an improvement. No papers reported the number of patients 'needed to treat' in order for one patient to benefit.

The majority of papers reported the mean QOL scores (77 (94%)) and 56 of these also indicated the variation around the mean (standard deviation, standard error or a confidence interval). Sixty-six (80%) papers reported p-values or an indication of the level of statistical significance of QOL differences. Eleven papers reported medians and six papers reported both means and medians. Three papers reported effect sizes. The summary measures used in the subgroup of high quality papers were very similar to the full set of papers.

3.4.5 Clinical significance and interpretation of results

Clinical significance was addressed in 31 (38%) papers (Table 3) and this was only marginally higher (44%) in the high quality papers. The most common method used was a change of >10 points to define a clinically relevant change (18 papers, 22% of all papers). This was usually referenced using Osoba *et al*(49), in which a change of 10-20 is "moderate". However, one paper referred to differences of 10 or more as large. Four other papers defined clinically meaningful change as any change from baseline, 5-10 points (not referenced), 8-10 points (Sloan(61)) and 10-15 points (Lee *et al*(62)) respectively, while three further papers defined different sizes as clinically meaningful depending on the scale. Two of these use a method used for the Uppsala questionnaire(63) and the other uses King's(19) estimates based on evidence-based effect sizes. Other methods of interpretation used were reference populations or norms (three papers) and effect sizes as defined by Cohen(31) or Osoba(49) (three papers). Two papers defined some results as clinically meaningful without defining the criteria used.

Clinical significance was not addressed in 51 (62%) papers; four of these papers contained no discussion of QOL differences and 47 (57% of all papers) relied mainly on statistical significance (or lack of statistical significance) in their discussion of whether there were changes in QOL (Table 4). These studies ranged in size from 48 to 791 patients (median 205 patients) and it is likely that at least some of these, generally large, studies will have found statistical differences in QOL that were too small to be of clinical relevance. For example, the largest study found differences in scores as small as 2.1 (physical and social functioning subscales) were statistically significant.

Table 4 Methods of interpretation

Clinical significance addressed	Methods used to assess size of QOL differences between groups or changes over time	Number of papers (% of total)
Yes*		31 (38%)
	Use of specified difference in score as clinically relevant:-	
	> 10 points	18 (22%)
	5-10 points	1 (1%)
	8-10 points	1 (1%)
	10-15 points	1 (1%)
	Subscale-specific	3 (4%)
	Comparison with reference population/norms	3 (4%)
	Effect sizes (<0.2 no change, 0.2-0.5 small, 0.5-0.8 moderate, >0.8 large)	3 (4%)
	Criteria for clinical significance undefined	2 (2%)
	Stable or improved from baseline defined as clinically relevant	1 (1%)
No		51 (62%)
	Statistical significance	47 (57%)
	No discussion of QOL differences	4 (5%)

*multiple methods used for 4 papers therefore numbers do not add to 31

3.5 Conclusion/Discussion

This review shows that RCTs using the EORTC QLQ-C30 report the QOL data to a high standard, with 70% meeting the criteria for high quality QOL outcomes. Similar reviews of QOL reporting in cancer trials have been carried out in prostate cancer(51;56), advanced breast cancer(64), colorectal cancer(65), non small-cell lung cancer(66) and, more recently, in complementary and alternative oncology medicine(67). These reviews generally show a lower standard of reporting, in particular the reporting of clinical significance ranged from 12% to 21% compared to the 38% seen here. This may be due to the earlier time period of studies included in previous reviews. Also, as our review was limited to studies using the EORTC QLQ-C30, three of the criteria were automatically met as the questionnaire is well validated, with

psychometric properties and cultural validity reported, and a range of QOL domains covered.

The main failings of the papers according to the checklist were not reporting the rationale for using the EORTC QLQ-C30 or the method of administration. These issues also arose in the previous reviews(51;56;64-67). In our review, this could be because the EORTC QLQ-C30 is well validated in cancer patients and authors referencing the validity of the questionnaire may regard this, implicitly, as their rationale for using it but without explicitly stating this they fail on this criterion. A more appropriate consideration regarding rationale for the chosen instrument may be whether QOL is relevant in the study at all, which QOL dimensions are important and therefore is the instrument chosen appropriate? The method of administration was only regarded as reported if the setting, e.g. questionnaires given in clinic or posted to patients at home, was reported. It was common to report that the questionnaire was self-reported but this was not considered sufficient to fulfil the criterion. An important aspect of administration is whether the assessments were before or after the clinical consultation, whether the results were confidential or whether they were used as part of the patient's management, which are details unlikely to be included in a paper.

A further finding of this review is that, perhaps not surprisingly, the RCTs meeting more of the criteria on the checklist were those with QOL stated as the primary outcome or reporting QOL as the main purpose of the paper. Papers reporting the full results of a clinical trial with QOL as a secondary outcome will have far less space to report the results therefore are likely to fail on more of the criteria. Papers reporting QOL alongside the main trial results are important if QOL is to have an impact on clinical decisions and on the results of clinical trials. Therefore it is unfair to penalise these papers because of the level of reporting of the QOL data when space in the manuscript will be limited. It is possible that there is a need for an even more minimal checklist in order that QOL results can be reported to a high standard despite limited space in the manuscript, otherwise such checklists may encourage separate reporting of the QOL results from the main trial results and could ultimately limit the overall impact of QOL data.

The majority of papers use a table or graph to display QOL results, which is encouraging as this is a good way of presenting QOL results from a number of subscales in a concise and clear way. Papers meeting the criteria for a high standard of reporting according to the checklist were more likely to use both tables and graphs to display the results. Surprisingly, three (5%) papers which met the criteria for high quality reporting used text alone to report QOL results. These papers reported p-values

for any significant results but little else. Although the checklist contains 'presentation of the results in general' as a criterion, this is based on whether the authors discuss the QOL outcomes giving any comments regardless of the results. This criterion can therefore be met by papers reporting very little QOL data resulting in them being classified as 'high quality' when they are clearly uninformative with regards to the QOL results. 18% of papers presented percentage of patients with improved/deteriorated/stable scores but the definition of an improvement varied. This is of concern, since papers which use any change in score as a 'clinically important change' and/or fail to specify a minimum time period will tend to overestimate the degree of change and therefore the impact of treatment.

Less than half of the papers addressed the clinical significance of QOL results. It is of some concern that half of the papers relied on statistical significance, or lack of statistical significance, rather than interpreting the magnitude of change *per se*. Whilst the minimum standard checklist for evaluating QOL outcomes(56) classes papers according to the robustness of QOL outcomes it is deficient in that it does not assess the appropriateness of the statistical analysis which is key to the QOL results. Given the apparent reliance on statistical significance in order to interpret results it is important that complex issues such as multiple outcomes, missing data and longitudinal data are dealt with appropriately in the statistical analysis and reported in suitable detail.

Where clinical interpretation was attempted, simple definitions were most common; generally >10 points was regarded as clinically significant. Osoba's work, however, was based on breast and lung cancer patients and the results may not be generalisable to other patient populations. The study was also carried out for specific subscales (global, physical, emotional and social functioning) and therefore these differences may not be applicable to other subscales. Although more detailed guidelines for interpretation are available for the EORTC QLQ-C30(19;31) a universal rule regardless of the subscale may be more attractive due to space limitations in manuscripts and the need for results to be easily understood, but if there are real differences in how to interpret the different subscales, then it is important to take this into account. A universal rule applied to all cancer sites and subscales may miss important differences in QOL or over-interpret differences that actually are of little clinical significance.

This review has highlighted that there are a number of methods of presentation and interpretation of QOL data available for use but that these are being applied regardless of their relevance to the specific QOL instrument or scale and could

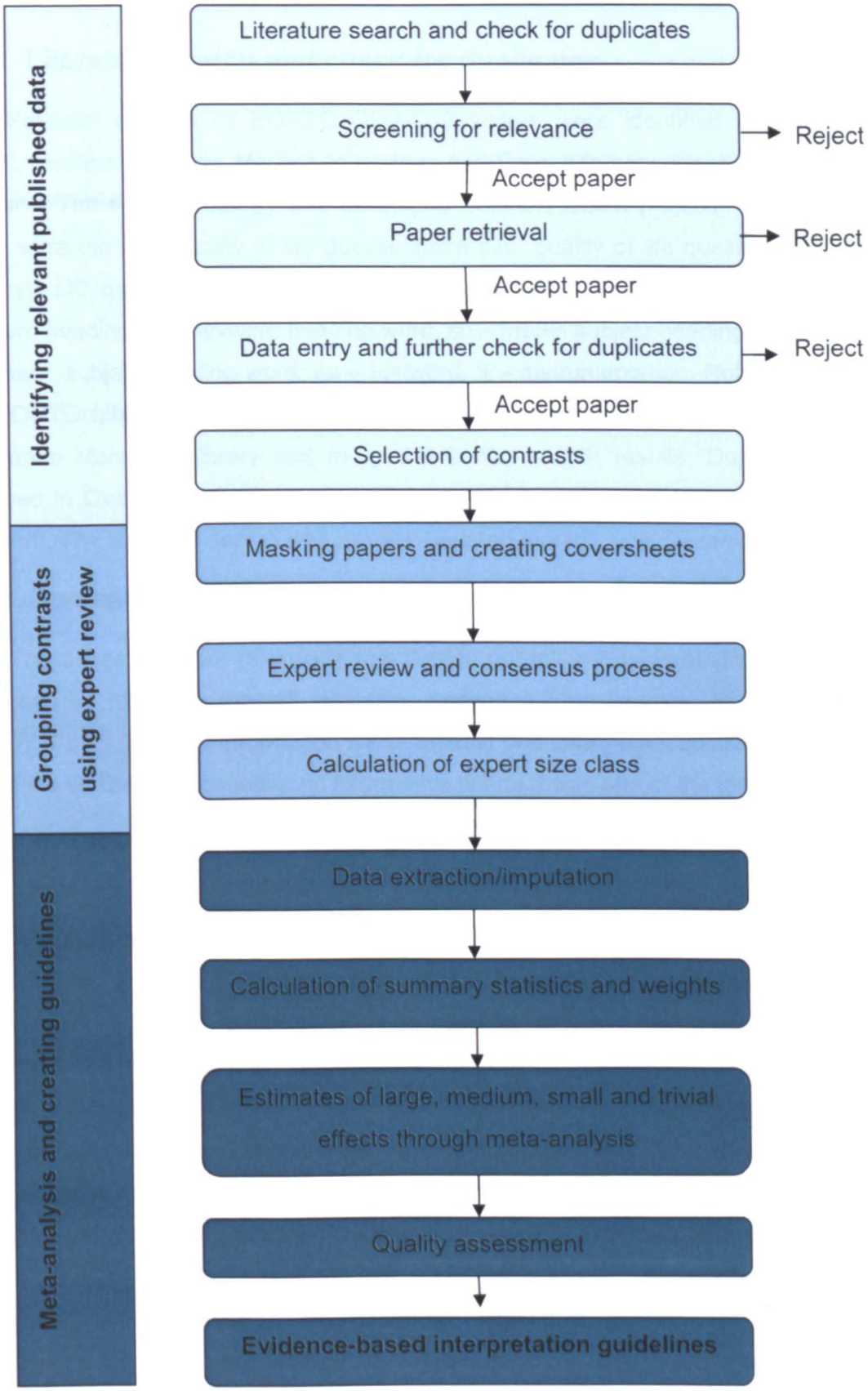
therefore be misleading. For accurate interpretation of QOL results, there is also a need to incorporate the adequacy of the statistical analyses in any assessment of the robustness of QOL outcomes. There is a need for further guidelines for the presentation and interpretation of results from the EORTC QLQ-C30 – and other QOL instruments – thus improving the ability of RCTs to influence treatment decisions.

4 Methods

4.1 Methods overview

Figure 5 shows an overview of the steps involved in producing the evidence-based interpretation guidelines. The project followed three phases. The initial phase involved searching for articles containing mean scores from the QLQ-C30. The second phase was the expert review process used to group the contrasts into trivial, small, medium and large. The final phase was the meta-analysis and derivation of the final guidelines. Each step is described in detail in the subsequent sections. In practice a number of the steps were overlapping, e.g. the literature search was updated monthly and new papers identified while the expert review process was on-going.

Figure 5 Overview of study processes



4.2 Identifying relevant published data

4.2.1 Literature search and check for duplicates

Potential sources of EORTC QLQ-C30 scores were identified by searching Cinahl, Medline, Embase, Medline-in-process and Psychinfo concurrently via the Ovid interface. The search strategy was developed from the EBES project(1). The search terms were qlq c30, quality of life questionnaire c30, quality of life questionnaire c30, eortc qlq-c30, qlq c33 and qlqc30+3. Fields searched were: ti – title, ab – abstract, kw – keyword heading, kf – keyword heading word, sh – MeSH subject heading, ot – original title, hw – subject heading word, tw – textword, it – instrumentation. References from the EORTC bibliography(60) supplemented the search. They were imported into a Reference Manager® library and merged with the search results. Duplicates were removed in Ovid or subsequently in Reference Manager by myself and the research assistant. The literature search was run and updated monthly until December 2006.

4.2.2 Screening for relevance

The assessment for relevance was carried out initially using abstracts and then confirmed by retrieving the pdf files (see Section 4.2.3). Sources were considered relevant if they provided information about at least one informative contrast for at least one of the QLQ-C30 subscales. An informative contrast was one of the following:-

- **Cross-sectional**: The mean difference between two independent groups.
- **Longitudinal**: The mean change within a group over time.

However, for the contrast to be truly informative it had to be based on a known anchor. Recall an anchor is defined as "an independent standard or anchor that is itself interpretable and at least moderately correlated with the instrument being explored."(12). Anchors were discussed in more detail in Section 2.1.1 and some specific examples of what we considered to be informative contrasts for the project follow in Table 5. Guyatt(12) stated that "the stronger the association, the more secure the inferences about interpretation of the target measure, and weak associations are liable to yield misleading results". I reviewed all anchors identified from the papers to ensure that the anchors we thought had the stronger and more well-known associations with QOL were prioritised for the review process. For example, we know there are small systematic effects of gender from the large reference population datasets and disease stage has been shown to separate groups of patients with respect to QOL during the validation of the QLQ-C30 questionnaire, therefore these were strong candidates for informative contrasts.

Table 5 Examples of informative contrasts and relevant anchors

Type of contrast	Anchor	Description
Cross-sectional	Gender	Males versus females
Cross-sectional	Treatment	Laparoscopic versus open surgery
Cross-sectional	Disease stage	Stage I versus Stage IV
Longitudinal	Time	Baseline versus 3 months
Longitudinal	Time	Diagnosis versus 1 year

The research assistant screened the abstracts according to the exclusion criteria listed in Table 6. I then reviewed any abstracts highlighted for potential exclusion. If further information was required I then retrieved the pdf file of the article and reviewed in detail against the exclusion criteria.

Table 6 Exclusion criteria

Exclusion criteria	Reasoning
Full paper unobtainable	Abstracts alone would not provide enough information for the project.
Non-English papers	Initially these papers were to be translated and included in the project where relevant. I developed a procedure for translations of non-English papers which involved an initial screening interview with the Research Assistant and translator in order to check through the exclusion criteria prior to full translation. The process was initiated but a larger number of papers requiring translations were identified than expected and the cost of translations was prohibitively high. This led to non-English papers being excluded.
Use of EORTC QLQ-C30 in diseases other than cancer	The EORTC QLQ-C30 questionnaire was developed specifically for cancer patients. There were however articles where the questionnaire was used in other disease areas. These were excluded as the questionnaire and the interpretation guidelines were aimed at cancer.
No QLQ-C30 scores reported	Articles such as narrative reviews and published protocols were identified in the literature search, which contained no actual data from the questionnaire.

Exclusion criteria	Reasoning
Mean EORTC QLQ-C30 scores not reported	Papers reporting only medians or correlations with EORTC QLQ-C30 data were excluded.
Only reporting the mean score of one group at one time (uninformative)	Papers had to contain at least one informative contrast (comparing across groups or within a group over time).
Overall sample size less than 10 or all contrasts contain a group with less than 10 patients	Although meta-analytic techniques can account for small studies by weighting in the analysis it was decided that data from 1-9 patients would not contribute greatly to the review. Generally these would be case studies from a single clinician or centre. The volume of work required to process these papers and obtain expert review outweighed the small contribution they would make to the interpretation guidelines. Further, they were likely to yield unreliable estimates.
Papers containing only contrasts of the following nature:- - comparing questionnaire administration method, e.g. computer vs paper or patient vs proxy - comparing languages, e.g. as part of a cross-cultural validation study - comparing cancer sites, e.g. breast vs lung, unless a comparison of cancer vs non-cancer/control.	These were excluded as they did not clearly represent a known clinical anchor on which to base estimates of size of difference.
Only the 'financial difficulties' subscale is reported	There is much debate over the inclusion of this subscale in a quality of life instrument and it was decided that papers which also had data from the other subscales should be prioritised.
No patient numbers reported	Required for the meta-analysis.

Exclusion criteria	Reasoning
Subscales not calculated according to EORTC scoring manual or use of an unknown score transformation	The EORTC questionnaire has a validated system for scoring into subscale scores. If this system was not used or if the scores were transformed from the 0-100 scale without explanation or details then they could not contribute to the meta-analysis.

I used Reference Manager® to manage the potential articles and wrote guidelines for the research assistant detailing the study processes described here. I wrote a specification for a Microsoft ACCESS database to store the data extracted for analysis and to track the progress of papers through the study. A programmer was used to build the database.

4.2.3 Paper retrieval

Following the initial screening the research assistant retrieved potential articles using on-line journals in the form of pdf files or photocopies from the library or document supply centre. The Reference Manager ID number was used as the link between the Microsoft ACCESS database, Reference Manager library and the paper or electronic copies of articles. I designed an EBIG Study Evaluation Sheet which was created for all papers (Figure 6). This contained the Reference Manager ID number along with the first author name, year, journal title and country. This sheet was produced for every article identified through the searches (except duplicates), including all rejected studies.

Figure 6 EBIG Study Evaluation Sheet

First Author and Initials		EndNote Ref No.	
Year	Journal		Country
		Entered on Database?	✓
		Accepted?	
		Rejected?	

REJECTED FROM:	✓	REASON:	✓
Abstract		No QLQ Scores Presented	
PDF		No mean scores presented	
Printed copy		Uninformative	
KC confirmed		Only Financial difficulties scale	
		Administration method	
		Validation study	
		Review	
		Other	

Notes:

Note that the full list of exclusion criteria in Table 6 was developed by coding the 'Other' category from this summary sheet as the project progressed. The rejection criteria on the evaluation sheet were those specified prior to starting the literature search. Note also that Endnote Ref No. refers to the Reference Manager ID number (the initial searches were managed using Endnote and later transferred to Reference Manager).

4.2.4 Data entry and further check for duplicates

A specifically designed data entry form was used to enter the details from the EBIG Study Evaluation Sheet. Rejected papers had the reason for rejection recorded and then were not processed any further. Where more than one reason was applicable

a hierarchy was used with the first exclusion criteria met in the above table being recorded.

Additional details were added to the database for papers accepted after this initial screening. These details included the type of study, source of the study sample, country of origin, language of QLQ-C30 questionnaire used, study name or number, type of primary cancer and extent of cancer. These fields were used to identify duplicate data, i.e. the same data reported in more than one source. Duplicate sources were identified by first looking at papers with authors in common. Papers reporting data from the same sample were still included in the project if they contrasted mean scores using different anchors. If papers reported on the same sample using the same anchor then the source providing the most detail about the sample and mean EORTC QLQ-C30 scores was included. Duplicate papers were retained for the review process if they provided additional clinical details of the study. For example, a clinical trial may have been reported in detail in one paper and the QOL results published separately. Although the latter provides the most QOL information it is unlikely to describe the study in as much detail as the main trial publication. In this scenario the papers were attached together for the expert review so the experts had more information on the study. Papers reporting the same study but with different anchors were also batched together for review.

Further basic data collection included details of the sample characteristics (age, sex, gender, disease site and disease stage), study design and study research question.

4.2.5 Selection of contrasts

Any QOL comparison from the paper (from a table, text or figure) was a potential contrast for the project. Therefore papers were likely to contain more than one informative contrast and the number could actually be quite high. For example, a single table reporting all 15 subscales of the QLQ-C30 for two treatment groups at baseline and one further time could lead to 60 contrasts (i.e baseline versus the second time within treatment group A, baseline versus the second time within treatment group B, comparison of treatment group A versus B at baseline and comparison of treatment group A versus B at the second time). Considering a paper may contain more than one table/graph containing QOL scores the number of contrasts could quickly become unmanageable and confusing to review. In order to avoid single papers dominating the project the number of contrasts was reduced according to defined rules. Contrasts were selected based on trying to obtain a dataset containing the full range of sizes of

QOL differences, i.e. we did not want to accidentally exclude very small or very large differences as all sizes were needed. If a number of points in time were reported then only three of them were taken forward (baseline, mid treatment and end of treatment or closest time to the end of treatment). If a number of cross-sectional contrasts were present various measures were used to reduce the number of possible comparisons. For example, if there were a number of treatment groups then one would be chosen as the 'control' group and only contrasts containing the control group used. If there were a large number of anchors then the strongest anchors were taken forward, i.e. those that are known to be closely linked to QOL. Papers with potentially large numbers of contrasts were discussed with the EBIG team for input, before I decided which contrasts to include on a case by case basis.

4.3 Grouping contrasts using expert review

In a traditional meta-analysis of randomised controlled trials (RCTs), studies may be grouped by dose level for example, in order to estimate the effect size within similar dose groups. Here we needed to group the contrasts into large, medium, small and trivial QOL differences in order to estimate effect sizes for each category. We sought to do this by using expert opinions on the appropriate size category for each contrast.

Expert opinion was sought from a panel of professionals with experience of cancer and the EORTC QLQ-C30. Following the literature review papers were categorised into different cancer areas. Where a paper covered more than one cancer they were usually categorised into an area of expertise instead e.g. radiology, chemotherapy, palliative care and so on. The expert panel was recruited based on the requirement for experts in all of the identified areas. The majority of experts were approached through the EORTC Quality of Life Group and in the UK the National Cancer Research Institute Clinical Studies Groups. The aim was to have at least three opinions for each paper so more reviewers were sought for disease areas with a large number of papers in order to reduce the workload for the project.

When a potential expert was identified they were sent an initial invitation letter. If they were willing to participate they returned details of their background and areas of expertise. Reviewers were then sent a manual containing an explanation of the project and the format of the review. The full expert review manual can be found at *www.ctru.leeds.ac.uk/ebig*(68) and the following sections describe the process in more detail. They were asked to return a practice example before reviewing any papers for

the project. Any issues with how the example was completed were raised by myself with the individual before they could commence reviewing.

The only papers excluded from the expert review were those that only contained a randomised contrast at the baseline time. These contrasts were assigned to the 'trivial' category automatically as randomised differences at baseline should be due to chance alone.

4.3.1 Masking papers and creation of coversheets

The aim was to obtain an assessment (blinded to the QOL results) of how experts believed the EORTC QLQ-C30 subscales would behave in each clinical situation defined in the contrast. We wanted a review from three experts for each paper in order to use an average opinion to group the contrasts in the meta-analysis. Experts were sent a copy of the paper along with a coversheet which showed the contrasts selected for review.

The reasoning behind obtaining a blinded assessment was to ensure that experts did not use pre-conceived ideas of the size of QOL differences to influence their ratings. For example, if an expert was familiar with Osoba's work(49) and regarded a difference of 10 points as a medium difference they could then use this as a benchmark if they could see the actual QOL scores in the original article. By blinding the experts to the QOL results they had to rely on their knowledge of the contrast and clinical setting and of using the QLQ-C30 to make a judgement on the likely size of QOL difference. Tables, graphs and text that contained QOL results (from both the QLQ-C30 and any other QOL questionnaire) were blacked out, or "masked". Also any interpretation of QOL results in the text was masked. Additional results from other QOL or patient-reported outcome (PRO) instruments were removed as these could give an indication of the QLQ-C30 results if they assessed similar aspects of QOL. The expert could therefore read about the type of study and setting, sample size, other results from the study, treatment details and adverse events if applicable.

The reviewer was asked to make a judgement on each contrast from the paper as to the relative size of differences (trivial, small, medium or large) they expected the domains in the QLQ-C30 to indicate for the chosen contrasts. If experts were familiar with the study they were asked not to try to recall the actual results from the study, but to judge what they might expect if they were conducting a similar study. If they felt they knew a study too well they could opt to return the paper without reviewing. They were asked to judge whether the QOL difference within a contrast would be a little, moderately or much better/worse. Definitions of large, medium, small and trivial were

provided in the manual as described in Table 7. The categories were given numbers; 0 for trivial, 1 for small, 2 for medium and 3 for large. A direction was also assigned to the scale so the experts were using a seven-point scale from -3 to 3.

Table 7 Definition of large, medium, small and trivial size categories

Size category	Description
Large	When circumstances have obvious and unequivocal clinical relevance (e.g. the contrast of patients with asymptomatic, early stage disease versus those with end-stage disease, or a treatment that is known to markedly improve the health state of most patients treated), group-level HRQOL is expected to be much better or worse, and large effects are expected.
Medium	When circumstances are likely to have clinical relevance, but to a lesser extent (e.g. for patients with metastatic disease, the contrast of those who respond to treatment compared those who do not respond, or a treatment that is known to be effective for a half of patients treated), group-level HRQOL is expected to be moderately better or worse, and moderate effects are expected.
Small	When effects are expected to be subtle but nevertheless clinically relevant (eg, the contrast of patients with regionally advanced cancer versus those with newly diagnosed metastatic disease, or a treatment that is known to improve the health state of only a small proportion of patients treated), group-level HRQOL is expected to be a little better or worse, and small effects are expected.
Trivial	When circumstances are unlikely to have any clinical relevance, group-level HRQOL is not expected to be any better or worse, and at best "trivial" effects are expected (including differences that may occur by chance). We use the phrase "much the same" for this size class.

A copy of the QLQ-C30 subscales and questions were also provided in the manual and experts were asked to keep these in mind while making their judgements, in order to consider what difference they expected clinically together with the difference the QLQ-C30 questions might detect.

Please refer to Appendix I for an example of a coversheet from the project. The coversheet gave brief details of the paper and the comparisons that they were being asked to make a judgement on. The brief details on the coversheet were intended to

summarise some key details of the paper, eg disease, disease extent, number of patients, type of study etc. and of the anchors used to contrast two groups of patients or a group over time. A blank summary table of the contrasts was then provided for experts to fill in with their ratings. The comparisons were divided into cross-sectional and longitudinal. Experts were advised to read the coversheet and then read the masked paper in order to understand the patient group and clinical setting for the comparisons. The tables only contained those subscales reported in the paper for each anchor. The numbers of patients in each group were included in the tables to give reviewers information regarding attrition in the study.

Preparing coversheets for the review was a time-consuming process which was open to human error in the transcription of information from the papers onto the coversheets. Initially the research assistant prepared the coversheets and then these were checked by myself or Prof Brown. In order to increase the efficiency of the process I wrote programs using SAS® software which could use the information directly from the Microsoft ACCESS database in order to populate the coversheets with the basic summary information from the study and also produce blank tables for the reviewers to fill in. This was output into Microsoft Word. Therefore the Research Assistant could simply enter the study details and contrast information into the database and then automatically produce the coversheet using the SAS® program. Although I still checked a proportion of the coversheets, the automation eliminated any possible error due to transcribing alone. Separate quality control checks were carried out on the database to check the information had been entered correctly so this also directly reduced any errors in the coversheets.

4.3.2 Expert review and consensus

4.3.2.1 Review process

Papers were allocated to relevant experts in batches, for example a single expert would be responsible for reviewing up to five papers in any four week period. Further papers were not sent to an expert until they had returned the previous batch of papers. Allocation was according to the area of expertise required and the availability of the individual. For some cancer sites there was a pool of experts to choose from, while for others (e.g. brain cancer) there were only two or three experts on the expert panel. The number of papers assigned to each reviewer was therefore determined by the number of papers identified in their area, their rate of returning papers after review and the length of time they were on the expert panel over the period of three years the review process was carried out.

4.3.2.2 Review format

Reviewers were asked to judge on a scale of -3 to 3 where 3 represented a large difference, 2 a medium difference, 1 a small difference and zero a trivial difference (Table 8). The negative or positive scale indicated the direction they thought the change would be in. For longitudinal contrasts a positive difference indicated an improvement in QOL scores over time and a negative difference represented a deterioration in QOL over time. For cross-sectional contrasts the direction simply referred to whether Group A was better than Group B or vice versa (see Section 4.4.1.2.2 for more details).

Reviewers were asked to use a percentage scale to indicate the certainty of their judgement. They could choose either one category or to spread their expectation over a few of the categories, with 0% representing completely uncertain, 100% representing completely certain. Reviewers also had the option of striking a line through, or returning a whole review, if they felt unable to make an assessment in that situation.

In the example in Table 8 the reviewer expected a small difference in physical functioning in favour of the younger group. They are not as certain for the cognitive functioning but indicate the difference could be trivial to small in favour of the younger group, with more emphasis on the small difference category. They are less sure again for the social functioning subscale, but indicate the difference should be trivial to small.

Table 8 Example of recording an expert reviewer's expectation

Anchor: Age Group 1: <60 years Group 2: >60 years	Group 1 worse post chemotherapy				Group 1 better post chemotherapy		
	-3	-2	-1	0	1	2	3
Physical Functioning					100%		
Cognitive functioning				25%	75%		
Social functioning				50%	50%		

4.3.2.3 Consensus

In the original EBES methodology(1;2) the experts chose one of the three size classes. Therefore it was clear to see where there were disagreements between the reviewers and initiate a consensus process. Because here reviewers could assign percentages to their ratings and rate in more than one of the size classes it was not as easy to see where major discrepancies lay. For that reason the only papers sent back to reviewers for querying were those where there was a discrepancy between

reviewers in the direction of the difference. The papers were returned to all reviewers for that paper to see if the direction had been confused when making their ratings. Reviewers were asked to check that their ratings were as intended for the specific contrasts where there was an issue. Reviewers either returned the coversheets after checking with any changes marked on (noting their reason for changing) or confirmed they were happy with the direction as originally marked. They did not enter into any discussions with the other experts, nor were they aware of the other experts' judgments.

4.3.3 Calculation of expert size class

Only papers with at least two reviews were taken forward for the analysis. For each reviewer, on each contrast, their weighted average was calculated using their percentage certainty as weights. So in the earlier example (Table 8) the weighted average for Cognitive Functioning from this reviewer would be (0.25×0)+(0.75×1)=0.75. This is referred to as an "individual opinion" on a contrast. The mean of these weighted averages from each reviewer on a contrast was then calculated, referred to as the "overall opinion" for that contrast. The overall opinion was then categorised into the small, medium and large categories for analysis using rounding, referred to as the "expert size class". i.e. if the overall opinion was 0.5 or greater but less than 1.5 then the expert size class was deemed to be 'small', greater than 1.5 but less than 2.5 would lead to a 'medium' size class and so on. The expert size class was used to group the contrasts in the meta-analysis in order to estimate the effect within each of the size groups. See Section 4.4.2 for the detailed meta-analysis methods.

4.4 Meta-analysis methods

4.4.1 Data extraction

4.4.1.1 Standard deviation extraction/derivation

I used standard deviations directly from the paper where possible. Where standard deviations were not provided, I calculated or estimated them using other information in the article where possible. If no information was available to estimate the standard deviation then I used imputation.

Derivation of standard deviations were based on the recommendations in the Cochrane handbook(69) and Follman(70), which are summarised in Table 9. For cross-sectional comparisons the pooled standard deviation was calculated where possible or

imputed. For longitudinal comparisons the standard deviation of the baseline time and the standard deviation of the change was calculated or imputed.

Table 9 Derivation of standard deviation (SD)

Data given	SD calculation
Means and standard errors for each group	$SD_{pooled} = \sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{n_1 + n_2 - 2}}$ $SD = SE \times \sqrt{n}$ Where SE is the standard error and SD is the standard deviation for each group.
Mean differences between groups and standard errors*	$SD = \frac{\text{SE of difference in means}}{\sqrt{\frac{1}{n1} + \frac{1}{n2}}}$
Means and confidence intervals for each group	NB Confidence intervals had to be symmetric around the mean in order to estimate SD $SD = \sqrt{n} \times (\text{upper limit} - \text{lower limit})/t$ Where t is the appropriate t-value for the significance level of the confidence interval and sample size (determined using Excel or t-tables).
Mean differences between groups and 95% confidence intervals*	$SE = (\text{upper limit} - \text{lower limit})/3.92$ $SD = \frac{\text{SE of difference in means}}{\sqrt{\frac{1}{n1} + \frac{1}{n2}}}$

Data given	SD calculation
Means for each group or difference between the means and p-value from t-test*	$SE = \frac{\text{difference in means}}{t}$ $SD = \frac{\text{SE of difference in means}}{\sqrt{\frac{1}{n1} + \frac{1}{n2}}}$ Where t was determined from the p-value using Excel or t-tables. For boundary p-values (e.g. p<0.05) the upper limit of the boundary was used in order to down-weight trials without full information. SD could not be estimated where a lower boundary alone was given (e.g. p>0.05).
Means/mean differences between groups and interquartile ranges	NB SD could only be estimated for large sample sizes (>60) with an approximate Normal distribution. SD could not be estimated for small sample sizes or for skewed distributions. $SD = Q75 - Q25 / 1.35$ Where Q75 and Q25 are the upper and lower quartiles.
Differences in change from baseline scores and SD from individual times	Where confidence intervals of change, t-values or p-values were also available then the above methods were applied. In the absence of any other information the SD had to be imputed instead.
Effect sizes (ES)	$SD = \frac{\text{difference in means}}{ES}$
Means/Mean differences between groups and ranges	Could not be estimated.
Means for each group and SD for whole group	SD_{pooled} could be estimated from SD of whole group(71) $SD_{pooled} = \sqrt{\frac{s^2(N-1) - \frac{(\overline{X}_{G1}^2 + \overline{X}_{G2}^2 - 2\overline{X}_{G1}\overline{X}_{G2})(n_{G1}n_{G2})}{n_{G1}+n_{G2}}}{N-1}}$

*In these situations there is insufficient information to derive the SD for each group or time. The calculated SD is the average of the two groups rather than the SD for each group. Using these derivations therefore assumes equal variances for the two groups.

If there was insufficient data available in the paper then I imputed standard deviations. Imputation was carried out assuming a common change variance(70) across contrasts for the same subscale, i.e. a weighted average of the reported change variances was used to impute for each subscale. This method of using other studies in the meta-analysis for imputation has previously been shown to provide accurate meta-analysis results(72). The level of imputation required in the dataset is summarised in Chapter 5.

4.4.1.2 Calculation of summary statistics for each contrast

I conducted separate analyses using two different outcome variables: mean differences and effect sizes. Mean differences were appropriate for the meta-analysis methods because all of the measurements from the papers were on the same scale(69), i.e. 0-100 as scored using the EORTC QLQ-C30 scoring manual(73). The standardized mean difference (or 'effect size') "is used as a summary statistic in meta-analysis when the studies all assess the same outcome but measure it in a variety of ways"(69), therefore is not necessary here where all studies are using the same outcome measure. The analysis of cross-sectional contrasts was however repeated using effect sizes as the outcome measure in order to compare our results to other guidelines for interpretation of QOL which commonly use effect size. Effect size was calculated as the mean difference divided by the best available estimate of between-person standard deviation.

Cross-sectional and longitudinal contrasts were analysed separately as they may be fundamentally different and warrant individual guidelines. Note that also the calculation of standardised mean differences (effect sizes) are fundamentally different for the between group comparisons and for comparisons over time (see Table 10 and Table 11). Standardised mean differences over time need to take into account correlation between observations on the same patients(71).

I weighted the contrasts in the meta-analysis using the inverse variance method. This has been shown to be the optimal way to combine estimates in a meta-analysis while accounting for the reliability of the estimates(74). If estimates were not weighted then contrasts with a small number of patients would be given equal weight to those with a large number of patients. The inverse variance method uses the sample size

and standard error to weight the contrasts in the analysis so larger studies are given more weighting.

4.4.1.2.1 Longitudinal contrasts

I calculated the mean differences for longitudinal contrasts so a positive difference indicated an improvement in QOL scores over time and a negative difference represented a deterioration in QOL over time for both symptom and function subscales. Since symptom subscales have a low score representing better QOL and function subscales have a high score representing the better QOL the calculations were reversed in order to achieve consistency in improvement or deteriorations over time across all contrasts. Table 10 details the calculations depending on which data were presented in the paper. Experts were asked to allocate a positive score if they thought the second time would be better than the first and a negative score if vice versa. They also did this regardless of the type of subscale so the calculation of mean difference and expert ratings were in the same direction. The standard errors (SE) presented in the table were used for weighting the contrasts in the analysis (see Section 4.4.2 for full details of the meta-analysis methods).

The notation used in Table 10 is as follows:-

$\bar{X}_{T1}$ = Mean score for baseline time

$\bar{X}_{T2}$ = Mean score for follow up time

SD_{T1} = Standard deviation of baseline scores

SD_{T2} = Standard deviation of follow up scores

SD_{pooled} = Pooled standard deviation

$\bar{X}_{T2-T1}$ = Mean change scores

SD_{change} = Standard deviation of the change over time

r = correlation between between baseline and follow up times, estimated at 0.5 if not given in the paper(75)

n_1= Number of patients at baseline time

n_2=Number of patients at follow up time

Table 10 Calculation of summary statistics and standard errors for longitudinal contrasts

Data available from paper or imputation	Mean difference (MD)	Definition of SD_{pooled}**	Standard error of MD	Effect size (ES)***	Standard error of ES
Means and sd for individual times where the subset of those with time 1 (T1) and time 2 (T2) are reported (i.e. $n_1=n_2$)	$\bar{X}_{T2} - \bar{X}_{T1}$	$\frac{SD_{T1}^2 + SD_{T2}^2}{2}$	$\sqrt{\frac{2SD_{pooled}^2(1-r)}{n}}$ Where n is the common sample size at time 1 and time 2	$\frac{\bar{X}_{T2} - \bar{X}_{T1}}{SD_{pooled}}$	$\sqrt{\frac{2(1-r)}{n} + \frac{ES^2}{2n}}$
Means and sd for individual times where all patients with T1 and all patients with T2 are reported (ie $n_1 \neq n_2$)	$\bar{X}_{T2} - \bar{X}_{T1}$	$\sqrt{\frac{(n_1-1)SD_1^2 + (n_2-1)SD_2^2}{n_1+n_2-2}}$	$\sqrt{\frac{2SD_{pooled}^2(1-r)}{n}}$ Where n is the average of the sample size at time 1 and time 2	$\frac{\bar{X}_{T2} - \bar{X}_{T1}}{SD_{pooled}}$	$\sqrt{\frac{2(1-r)}{n} + \frac{ES^2}{2n}}$
Mean change and sd of change	$\bar{X}_{T2-T1}$	$\frac{SD_{change}}{\sqrt{2(1-r)}}$	$\sqrt{\frac{SD_{change}^2}{n}}$	$\frac{\bar{X}_{T2-T1}}{SD_{pooled}}$	$\sqrt{\frac{2(1-r)}{n} + \frac{ES^2}{2n}}$

NB These calculations are for the functional subscales. For symptom subscales, a multiplier of -1 was used. Calculations were sourced from Lipsey and Wilson(75), Follman(76) and The Cochrane Handbook(77). Some specific scenarios were derived.

**May also be derived using Table 9 where needed.

***This effect size estimate has been shown to be upwardly biased, particularly for small sample sizes(78), therefore as recommended by Lipsey and Wilson(75) I have subsequently applied a correction of $\left[1 - \frac{3}{4N-9}\right]$, where N is the total sample size for the contrast, to each effect size.

4.4.1.2.2 Cross-sectional contrasts

For cross-sectional contrasts a negative or positive mean difference had less meaning. The order of the two groups for comparison was chosen as consistently as possible, for example, Group 1 was always the control group in a treatment study, the lowest age group, the less advanced disease stage etc. This attached similar meaning or 'direction' to similar contrasts within the same anchor. However, for some cross-sectional contrasts there was not an obvious choice of 'control' group, for example outpatient versus inpatient contrasts or a comparison of two commonly used therapies. In these cases the direction of the difference was more arbitrary and the groups were allocated in the order they appeared in the original article.

As for the longitudinal contrasts the calculation was reversed for the symptom subscales so that a positive difference on a function subscale had the same meaning as on a symptom subscale. Experts were asked to give a positive score if they thought Group 2 would be better than Group 1 and negative if they thought Group 2 would be worse than Group 1. The mean differences were calculated as shown in Table 11 in order to match the direction the experts were judging in. The calculations are shown for functional subscales. The calculations for symptom subscales were the same with a multiplier of -1. The standard errors (SE) presented in the table were used for weighting the contrasts in the analysis (see Section 4.4.2 for more detail). Note that some papers presented change from baseline scores rather than mean scores for the groups therefore the difference between groups was the difference in these change scores.

The notation used in Table 11 is as follows:-

$\bar{X}_1$ = Mean score for Group 1 (or Phase 1 for a cross-over study)

$\bar{X}_2$ = Mean score for Group 2 (or Phase 2 for a cross-over study)

n_1 = Number of patients in Group 1 (or Phase 1 for a cross-over study)

n_2 = Number of patients in Group 2 (or Phase 2 for a cross-over study)

N = Number of patients in the cross-over trial

SD_{pooled} = Pooled standard deviation

$\bar{X}_{AB}$ = Mean score for Group *A* at time *B*

SD_{change} = Standard deviation of the difference between two measurements on an individual

r = Correlation between phase 1 and phase 2 scores (for cross-over studies) or between time 1 and time 2 (for change from baseline), estimated at 0.5 if not given in the paper(75)

MD= Mean Difference

Table 11 Calculation of summary statistics and standard errors for cross-sectional analysis

Data available from paper or imputation	Mean difference (MD)	Definition of SD_{pooled}**	Standard error of MD	Effect size (ES)***	Standard error of ES
Means and standard deviations for each group	$\bar{X}_2 - \bar{X}_1$	$\sqrt{\frac{(n_1-1)SD_1^2 + (n_2-1)SD_2^2}{n_1+n_2-2}}$	$SD_{pooled}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$	MD/SD_{pooled}	$\sqrt{\frac{n_1+n_2}{n_1n_2}+\frac{ES^2}{2(n_1+n_2)}}$
Mean change from baseline for each group and standard deviation of the change	$(\bar{X}_{22}-\bar{X}_{21}) - (\bar{X}_{12}-\bar{X}_{11})$	$\frac{SD_{change}}{\sqrt{2(1-r)}}$	$SD_{change}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$	MD/SD_{pooled}	$\sqrt{\frac{n_1+n_2}{n_1n_2}+\frac{ES^2}{2(n_1+n_2)}}$
Means and standard deviations for each phase (cross-over study)	$\bar{X}_2 - \bar{X}_1$	SD_{pooled} is the SD of a single measurement or pooled SD from phase 1 and phase 2 if cannot assume both are equal. i.e. $\sqrt{\frac{(n_1-1)SD_1^2 + (n_2-1)SD_2^2}{n_1+n_2-2}}$	$SD_{change}\sqrt{\frac{1}{N}}$ where $SD_{change} = \sqrt{2\times(SD_{pooled})^2(1-r)}$	MD/SD_{pooled}	$\sqrt{\frac{1}{N}+\frac{ES^2}{2N}} \times \sqrt{2(1-r)}$

NB These calculations are for the functional subscales. For symptom subscales, a multiplier of -1 was used. Calculations were derived from a combination of those given in Lipsey and Wilson(71), Follman(70) and The Cochrane Handbook(77)

**May also be derived using Table 9 where needed.

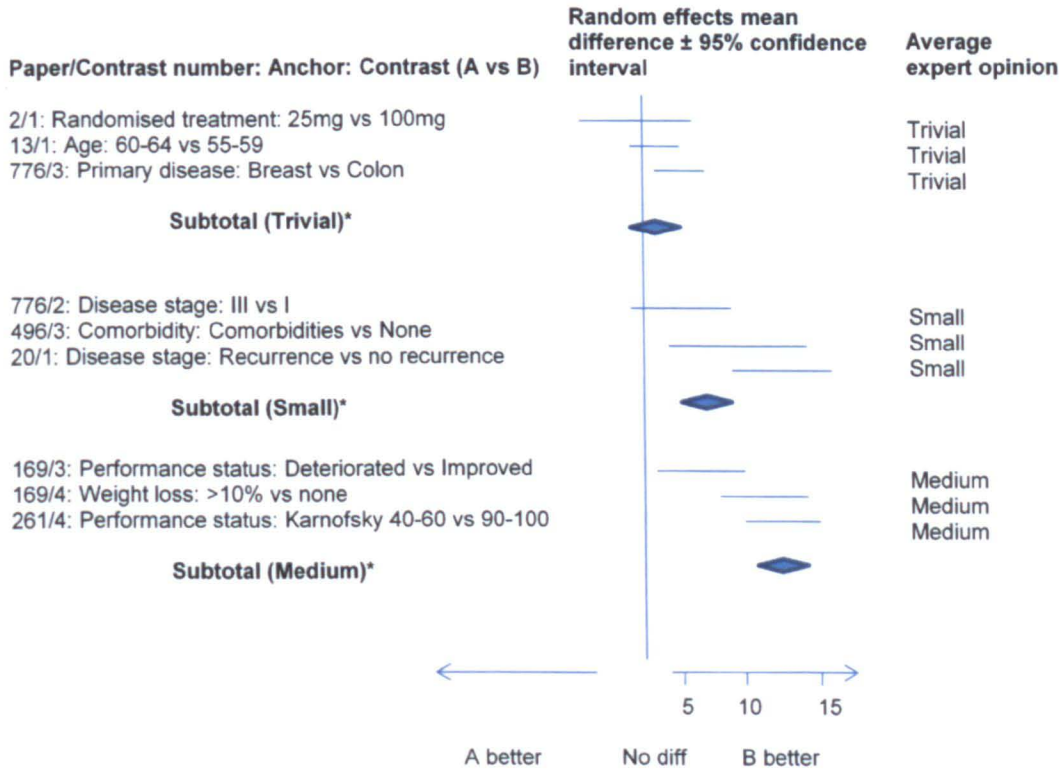
***This effect size estimate has been shown to be upwardly biased, particularly for small sample sizes(78), therefore as recommended by Lipsey and Wilson(71) I have subsequently applied a correction of $\left[1-\frac{3}{4N-9}\right]$, where N is the total sample size for the contrast, to each effect size.

4.4.2 Estimating large, medium, small and trivial effects using meta-analysis

4.4.2.1 Meta-analysis approach

The meta-analysis was carried out using SAS® Software. I wrote macros to implement the analysis by adapting those written by Wang and Bushman(79). I conducted meta-analyses for each subscale separately in order to create guidelines for specific subscales. Hence with 15 subscales and two outcome variables (mean difference and effect size) there were 30 models constructed to combine the cross-sectional contrasts and 30 for the longitudinal contrasts. The contrasts were grouped in the meta-analysis by the expert size class in order to obtain an estimate for large, medium, small and trivial effects. This is akin to a standard meta-analysis of RCTs where studies may be grouped by dose, for example, in order to estimate the effect size within each group of studies using similar doses. The following Forest plot (Figure 7) is used to illustrate how the analysis estimated an effect size for trivial, small, medium and large differences. Note that plots of this nature have not been used to display the actual results due to the large number of contrasts that would be required in each figure.

Figure 7 Forest plot to illustrate meta-analysis methods



4.4.2.2 Fixed effect models, tests for heterogeneity and meta-regression

Initially I planned to carry out the analysis using a fixed effect model (weighting contrasts using the inverse variance method). This was the method used for the EBES project. A fixed effect model for meta-analysis assumes that each observed effect estimates an overall population effect (with a random error only due to chance). Here this meant assuming that within each subscale and for each size class the observed effect for the contrast was the common effect plus a within-contrast error, i.e.

$$Y_{ijk} = \beta_0 + \gamma_{ijk}$$

Where Y_{ijk} is the mean difference or effect size for the i^{th} contrast in size class j and subscale k, β_0 is the population effect for that subscale and size class and $\sigma^2_{Y_{ijk}}$ is the variance due to sampling error for the i^{th} contrast.

The weighted mean ($\bar{Y}_{ijk}$) within each expert size class (j) and for each subscale (k) was calculated using the inverse-variance as weights:-

$$\bar{Y}_{ijk} = \frac{\sum C_{ijk}/SE^2_{ijk}}{\sum 1/SE^2_{ijk}}$$

Where C_{ijk} = the effect estimate (mean difference or standardised mean difference) for the i^{th} contrast in size class j and subscale k and SE_{ijk} was the standard error for that estimate. These were extracted from the data as described in Table 10 and Table 11.

Fixed effect meta-analysis ignores between-contrast variation. As a result, parameter estimates are biased if between-contrast variation cannot be ignored. In a traditional meta-analysis very similar studies are usually combined, to compare Drug A vs Drug B for example, and this assumption may therefore be appropriate. However, here, the contrasts being combined were not only from different studies across different cancers but they also covered a wide range of anchors. Further, even within an anchor the contrasts could vary greatly in what they were comparing. Therefore I carried out tests for heterogeneity(77) to indicate if the fixed effects models were appropriate in this situation. The Q-statistic (a weighted sum of squared deviations from the mean) was calculated,

$$Q = \sum (1/SE^2_{ijk}) \times (C_{ijk} - \bar{Y}_{ijk})^2$$

The Q-statistic has a chi-square distribution with degrees of freedom n-1 where n is the number of contrasts for that subscale and size class. If the value of Q is larger than would be expected from the chi-square distribution then the hypothesis of a single population mean would be rejected.

The tests for heterogeneity were highly significant for all models here indicating that the variation between contrasts was greater than would be expected by chance alone. I then used meta-regression to explore if the extra variation could be explained by characteristics of the studies/contrasts. I identified a group of potential factors for the meta-regression that could explain the heterogeneity between the contrasts. These were discussed and agreed with the rest of the team prior to any analysis. Table 12 shows the list of factors explored.

Table 12 Factors for meta-regression

Applicable to which contrasts	**Paper or contrast level factor**	**Factor**	**Description**
Both cross-sectional and longitudinal	Contrast	Expert size class	Trivial, Small, Medium, Large
Both cross-sectional and longitudinal	Paper	Design	Cohort, Multiple (contains multiple studies), Non-randomised phase I, Non-randomised phase II, RCT Phase II, RCT Phase III, RCT (phase not specified).
Both cross-sectional and longitudinal	Paper	Disease	Brain, Breast, Colorectal, Gastro-Intestinal, Gynaecological, Haematological, Head and neck, Lung, Mixed, Prostate, Testicular, Urology/Kidney
Cross-sectional	Contrast	Anchor	Disease related, Treatment/Intervention/Assessment related, Patient characteristics related, Physical function related, Symptom/Emotional related, Time related, Survival related

Applicable to which contrasts	Paper or contrast level factor	Factor	Description
Cross-sectional	Contrast	Timing of the contrast	Baseline, Other, Not specified
Longitudinal	Contrast	Timing of second assessment	Time in months if known
Longitudinal	Contrast	Dropout	Proportion of patients dropping out from Time 1 to Time 2 for each contrast. Note this is zero if the subset of patients with both times is presented. NB this does not necessarily represent attrition across the study more generally.

Initially the expert size class was added to the model as the grouping variable in order to show whether accounting for the expert size class alone explained the heterogeneity between the contrasts. Then each of the other factors was added individually to see if heterogeneity remained after accounting for the expert size class and one extra factor. Multivariate meta-regression analyses were not carried out due to the probable collinearity between the proposed factors. The meta-regression model can be written as follows:-

$$Y_{ik} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \gamma_{ik}$$

Where Y_{ik} is the mean difference or effect size for the i^{th} contrast in subscale k, β_0 is the population effect for that subscale, x_1 represents the size class (with coefficient β_1) and $x_{2\ is}$ one of the factors from Table 12 (with a coefficient of β_2) and $\sigma^2_{\gamma_{ik}}$ is the variance due to sampling error for the i^{th} contrast.

None of the factors adequately explained the heterogeneity. Following the advice in Wang and Bushman(79) I concluded that a random effects model was more appropriate since the variation in contrast effects could not be explained by the expert size class and any one additional factor.

The guidelines were therefore derived from random effects models. The results from the fixed effect models are not reported in the thesis as the model was clearly inadequate for the wide range of contrasts being combined.

4.4.2.3 Random effects model

A random effects model assumes the population effect to be estimated is actually a distribution rather than a single value. The size of the effect has its own mean and variance. This assumption behind the random effects model seemed reasonable for the kinds of contrasts being combined here. We assumed that within the size classes the contrasts lead to similar but not identical effect sizes. The basic model for the random effects model is as follows:-

$$Y_{ijk} = \beta_i + \gamma_{ijk}$$

Where Y_{ijk} is the mean difference or effect size for the i^{th} contrast in size class j and subscale k, β_i is the local population effect for that subscale and size class and $\sigma^2_{Y_{ijk}}$ is the variance due to sampling error for the i^{th} contrast.

Models fitted the expert size class as a fixed effect and contrasts nested within papers as a random effect. Contrasts were nested within papers in order to account for multiple contrasts from each paper. Expert size class was used in the model to group the contrasts in order to estimate effects for trivial, small, medium and large separately. The effect size for each contrast was weighted by the inverse of the variance of the estimate for that contrast plus an estimate of the additional random variation between contrasts in that size class,

$$1/\{SE_i^2 + \hat{\sigma}_\theta^2\}$$

Where SE_i^2 was the estimator of the variance of the outcome variable (either mean difference or effect size) from the i^{th} contrast within the size class and $\hat{\sigma}_\theta^2$ was the estimator of the random effects variance for that size class. The random effects model variances were estimated using the residual (restricted) maximum likelihood (REML) method.

4.4.3 Quality assessment

4.4.3.1 Background

4.4.3.1.1 Quality assessment in standard meta-analyses

In the standard application of meta-analysis, quality assessment means investigating criteria that may affect the validity of the meta-analysis results, i.e. looking at each study and assessing the extent to which a study's design and conduct are likely to prevent systematic errors or bias(69). A number of scales or checklists have emerged that assess the methodological quality of clinical trials(80), largely focussing on whether a trial is randomised or conducted in a blinded fashion. Once a quality assessment has been carried out there are a number of options for incorporating quality assessment into the meta-analysis(69;81); from ignoring quality, weighting according to quality through to excluding poor quality studies. Each approach and problems with the approaches are discussed below:-

- Ignore the variation in quality. The reasoning behind this approach is that the quality assessment is subjective and a meta-analysis should be objective. However, when you consider an individual study the degree to which you believe in the results comes from an assessment of the study and its conduct. So when considering a meta-analysis it would seem at odds to disregard any indication of study quality.
- Use exclusion criteria to exclude studies with a high risk of bias. This effectively then ignores any variation in quality of the included studies and the cut-off point for inclusion is arbitrary.
- Use measures of validity or quality to explore the data via subgroup analyses or sensitivity analyses. For example, arranging meta-analysis plots of results in order of validity in order to show where the differences lie or analysing subgroups above and below cut-points for validity. This gives more insight as to how validity effects the outcome measure but does not make a decision on how to treat the different subgroups in the analysis.
- Use quality assessment to weight studies in the analysis or to explore the magnitude of effects across studies using meta-regression. Weighting is generally not recommended as it is fairly arbitrary as to how to convert quality scores into weights. The Cochrane Collaboration strongly discourage the use of quality weighting scales(77)

In practice a combination of these methods is useful rather than considering one approach to the issue of quality.

4.4.3.1.2 Meaning of quality for EBIG

This was not a standard application of meta-analytic techniques and the meaning of quality for this analysis was likely to differ from that in standard meta-analyses. The units being combined were not studies, which have a widely accepted hierarchy for the quality of their design and reporting (see for example *www.ebmpyramid.org*(82)). This pyramid treats systematic reviews as the highest, or best, level of evidence and randomised controlled trials as the best evidence from individual studies. However, here the units being combined are contrasts rather than studies. Although study design may have a bearing on the quality of the resulting contrasts from that study it is not clear that all contrasts from the same study would thus be of the same quality.

The standard quality criteria were derived with treatment comparisons in mind. The RCT would therefore provide the gold standard of evidence because the randomisation should ensure the characteristics of the treatment groups were very similar, in terms of both observed and unobserved characteristics. A cohort study does not have the bias protection of randomisation, and therefore would be considered evidence of a lower quality. While the notion that study designs have a certain hierarchy of evidence may apply directly to the contrasts in our meta-analysis which compare treatments, this makes up only a small proportion of the contrasts included in the analysis.

For other contrasts the standard measures of quality may not be relevant. Say we were combining a contrast comparing males versus females from a RCT with the same contrast from a cohort study. The groups of males and females are a product of the sample in the study and not affected by whether the study is randomised. Therefore it is not as clear that the RCT provides a higher standard of evidence for our meta-analysis. In fact, the RCT may have a study population more refined by other inclusion and exclusion criteria than the standard population of cancer patients. This may make decisions on the size of possible QOL differences more difficult for an expert to predict than if the contrast was from a cohort study with a population more like they are used to seeing through the clinics.

Despite the differences between our analysis and standard meta-analyses the aim of investigating quality however was the same; to explore factors that could affect the validity of the meta-analysis results. The aim of quality assessment was to look for factors that modified the effect size in some way. In our methodology we considered that there were two areas which could affect the quality of the resulting interpretation

guidelines; the quality of the studies (paper level factors such as study design and attrition) and the quality of the expert reviews (contrast level factors). The quality of the study could affect the size of the QOL differences observed in the study, therefore directly impacting on the effect size for analysis. The quality of the experts' reviews could also impact the effect size since if the expert size class was incorrect then the contrasts would be pooled inappropriately and the resulting size class estimates would be incorrect.

The size class which grouped the contrasts for analysis was obtained using the average expert judgement. Factors that could affect the reliability of the average score from experts were therefore important. Firstly, there was the issue of agreement between reviewers and how this affected the average review score. If a contrast had two reviewers who disagreed in their judgements then taking the average would place the contrast in a size class that was not actually intended by either reviewer. Secondly we considered the uncertainty around the expert ratings. If there was a large amount of uncertainty in where the reviewers placed their judgments then using the average to define size class may also be fairly meaningless.

We also considered that there may be factors that made it harder for experts to predict the QOL differences. The contrasts the experts found hard to judge may not be as useful for deriving the guidelines. Qualitative interviews with the experts (described in Section 4.4.3.2) were used to identify any possible factors making their review more difficult so these could be explored in relation to quality.

4.4.3.1.3 Approach to incorporating quality assessment in EBIG

Since the factors which may affect quality were largely unknown in this application of meta-analysis I considered that weighting the analysis by a derived quality score would be inappropriate. This method is already criticised as being too arbitrary and here it would have been even harder to attach a score to the various factors we identified as possibly affecting quality.

In a traditional meta-analysis it is important to include all of the available evidence to avoid bias. Here we have an added level of complexity in that the grouping variable, the expert size class, is a subjective measure. The expert size class was also derived in an experimental way so we could not be sure of its accuracy and reliability up front for each contrast. Rather than trying to synthesize all of the available evidence I considered it more appropriate to identify the contrasts where the expert size class was

most likely to be accurate and then exclude the contrasts of a poor quality when deriving the guidelines. That way only the 'best' evidence contributed to deriving the guidelines and more faith could be put in the results.

4.4.3.2 Interviews with experts

All members of the expert panel were invited to take part in a short telephone interview in order to provide feedback on the review process and to obtain details of their approach to the reviews. Interviews were conducted when reviewers had completed the review of some papers but, where possible, were still involved in the review process. Full details of the interview questions are provided alongside the results in Chapter 6.

The expert review process was a new approach developed from the original methods used for the EBES project. The EBES expert panel consisted of only three experts (clinical oncologists) who reviewed all papers regardless of the disease area. For our project papers were allocated to different reviewers from a larger panel of reviewers, according to either the disease area or the specific treatment or other specialist area if papers contained a mixture of disease areas. Although the instruction manual(68) was written to try to ensure a consistent approach between reviewers it was unknown whether a larger panel, with different backgrounds, would approach the reviews consistently.

The quality of the interpretation guidelines directly relies on the quality of the expert reviews, since the estimates from the meta-analysis are grouped using the expert opinion. Interviews with the experts were used to identify factors that could affect the quality of the reviews, such as aspects the experts found difficult and how they approached the task. These factors were then explored quantitatively by investigating their effect on the experts' uncertainty and concordance between the expert reviews, described in the following sections.

4.4.3.3 Meta-analysis quality

4.4.3.3.1 Exclusion of small studies and contrasts

We made a decision at the start of the project to exclude any studies containing less than ten patients. These were likely to be the poorer quality studies and would provide unreliable estimates of mean change due to the small sample size. During the

project it also became apparent that, due to the same arguments, individual contrasts with any group containing less than ten patients should also be excluded.

4.4.3.3.2 Data checking and cleaning

In order to ensure the best quality data there were a number of data checking processes used during the project. A proportion of the data entered on the database (both mean differences from the original papers and expert review scores) was reviewed by myself and Prof Brown to check for data entry errors. Initially 100% of the data entered was checked and once the error rate was sufficiently low only a proportion were subsequently checked. I also wrote a data cleaning program in SAS® to highlight any reviews where the reviewers weighted average disagreed by more than two size classes. This was run monthly by the research assistant. Where discrepancies were found the data entry of the reviewers scores were checked first and then, assuming there were no errors in the data entry, the paper and coversheets were sent back to each reviewer on that paper to check the review was as they had intended. The contrast with a discrepancy was highlighted so the reviewers only had to check the contrast in question. However, the reviewers still had no knowledge of the nature of the discrepancy or of the other reviewers' scores.

4.4.3.4 Expert review quality

The following three sections explain the methods used to measure correlation between the experts and the observed QOL differences, concordance between the reviewers and uncertainty in their ratings. These analyses were only relevant for the subset of contrasts that underwent the expert review. Therefore the contrasts we assigned to trivial (as they were baseline contrasts between randomised groups) were excluded from these analyses.

4.4.3.4.1 Correlation with mean differences in original paper

Correlation was used to assess the degree to which reviewers judgements were reflecting the actual score differences in the papers. Non-parametric (Spearman's rank) and parametric (Pearson's) measures of correlation were both calculated along with p-values at the 5% level to test the null hypothesis for each reviewer that their correlation was equal to zero, i.e. no correlation between their ratings and the observed scores.

4.4.3.4.2 Concordance between reviewers

The 'agreement' or 'concordance' between different reviewers for the same contrast was difficult to define because experts could assign percentages across the 7-point scale to represent their view of the size in QOL change. Although measures of agreement have been developed for ordinal scales similar to our rating scale, the measures I found through searching the literature all assume that raters were placing their ratings in only one of the categories, rather than potentially across a number of categories. Because there was a lack of a standard measure for this particular type of data I explored a number of possible measures.

Two measures were used which required each reviewer's weighted average on a contrast to be rounded up or down to the closest size class. The first measure looked at the distance between the reviewers in terms of number of size classes apart(1) and the second method used a consensus measure derived using information theory which gives a score between 0 (complete disagreement) and 1 (complete agreement) (83). Although these provide simple summaries there is a loss of information when rounding the weighted averages up or down. For example, a contrast with two reviews with weighted averages of 1.4 and 1.6 would have the same agreement as a contrast with two reviews further apart such as 1.0 and 2.0. Two further measures were therefore used which were based on the continuous weighted average measure rather than requiring any rounding into categories; Intra-Cluster Correlation (ICC) coefficients(84) and the between-reviewer standard deviation. The details for these four methods are as follows.

1) After rounding each reviewer's weighted average back onto the original scale (-3 to 3 as described in Section 4.3.1) the maximum distance between reviewers on the same contrast was calculated. The proportion of contrasts with exact agreement, a distance of one category apart, two categories apart and so on were calculated. These are displayed in pie chart and provide an easy visual summary of how closely the experts agreed.
2) The consensus measure (Cns) was calculated for each contrast as follows(83);

$$Cns(X) = 1 + \sum_{i=1}^{n} p_i log_2\left(1 - \frac{|X_i - \mu_x|}{d_x}\right)$$

Where X represents the scale used by the reviewers (-3, -2, -1, 0, 1, 2, 3), X_i is the i^{th} member of that scale, p_i is the proportion of reviewers with their rounded weighted average in the i^{th} category for that contrast, d_x is the width of the scale (i.e. 7) and μ_x is the mean score. The calculation is illustrated for a range of consensus values in Table 13.

Table 13 Calculation of consensus between reviewers

Contrast (Paperid: Contrast: Subscale)	Reviewers weighted averages	Number of reviewers in each category after rounding							Mean (μ_x)	Cns
		-3	-2	-1	0	1	2	3		
13:5:PA	-3,1.8	1	0	0	0	0	1	0	-0.50	0.36
785:2:PF	-1.5,-3.0,2.5	1	1	0	0	0	0	1	-0.67	0.35
13:12:AP	0.4,2.5	0	0	0	1	0	0	1	1.50	0.65
916:2:EF	0,1.5,0	0	0	0	2	0	1	0	0.67	0.80
23:1:QL	-2,-1.9,-1	0	2	1	0	0	0	0	-1.67	0.90
17:2:PF	0.8,1.0,0.5	0	0	0	0	3	0	0	1.00	1.00

3) The ICC measured the degree of clustering between the weighted average of each reviewer on the same contrast. The ICC was calculated for each subscale and for longitudinal and cross-sectional contrasts separately. If $X_{ij} = \mu + B_j + W_{ij}$, where X_{ij} is the i^{th} rating (weighted average for each reviewer) on the j^{th} contrast. Mu is the overall population mean of the ratings, B_j is the difference from μ of the j^{th} contrasts 'true score' (i.e. the mean across repeated ratings on the j^{th} contrast) and W_{ij} is the residual component (containing inseparable effects of the reviewer, reviewer by target interaction and error term). If the residual variance is σ_r and the variance of B_j is σ_b then the ICC is defined as $\sigma_b/(\sigma_b+\sigma_r)$ or, in words, the proportion of between variance in total. The meaning of ICCs and how sensitive they would be to differences between reviewers however was unclear, since the original scale used by the reviewers was a 7-point categorical scale and the calculated weighted average was then used as the outcome variable. As a rule of thumb ICCs of <0.4 were defined as poor agreement, 0.4 to 0.75 as fair to good agreement and >0.75 as excellent agreement(85).

4) The standard deviation between the weighted means from reviewers judging the same contrast was calculated as an alternative measure of concordance. The standard deviation ($SD_{between}$) was calculated for each contrast using the distance of the weighted mean for each reviewer from the overall mean of the two or three reviewers. For example, for a single contrast with r reviewers and a mean weighted average score of $\bar{\mu}$,

$$SD_{between} = \sqrt{\frac{1}{r-1}\sum_{i=1}^{r}(\mu_i - \bar{\mu})^2}$$

4.4.3.4.3 Uncertainty in reviewers scores

Uncertainty within a reviewer's judgment could also be defined in a number of ways. Simplistically it could just be the number of categories each reviewer places their percentages in, i.e. if a reviewer rated 100% in one category they are certain of their judgement whereas if they marked percentages across three categories they are less certain. However, this would rate uncertainty as the same for a reviewer with judgements in two categories split 90%:10% compared with a review split 50%:50% from a different reviewer. It is clear therefore that the weight in each category is important as well as the number of categories used. Summaries of the average number of categories used for a contrast and the average peak (highest weight) used are easy to interpret and therefore were used as simple summaries of the uncertainty.

A different approach to uncertainty was used for the purpose of exploring factors affecting uncertainty. Uncertainty was defined by calculating the standard deviation for each review from its weighted mean. If w_1 to w_7 were the weights assigned by a reviewer to each of the categories on the -3 to 3 scale (divided by 100 so they add to 1) the weighted mean for that review was calculated as

$$\mu = \left[(w_1 \times -3) + (w_2 \times -2) + (w_3 \times -1) + (w_4 \times 0) + (w_5 \times 1) + (w_6 \times 2) + (w_7 \times 3)\right]$$

The standard deviation within a review was calculated using the formula for calculating the standard deviation of a weighted mean.

$$\sigma_{weighted} = \sqrt{\frac{\left(\sum_i w_i (x_i - \mu)^2\right)}{(N-1)\sum_i w_i / N}}$$

where the x_i are the -3 to 3 on the rating scale, N is the number of non-zero weights.

Since the sum of our weights equates to 1 this is equivalent to

$$sd_w = \sqrt{\frac{N}{N-1} \times \sum_i w_i (x_i - \mu)^2}$$

This is referred to as the within-reviewer standard deviation. The average of the within-reviewer standard deviations for each contrast was used as the measure of uncertainty.

4.4.3.4.4 Investigation of factors possibly affecting the quality of expert reviews

Factors related to overall study quality and other factors highlighted from the expert interviews as potential issues with the review process were investigated to see which influenced the quality of the expert review. The standard deviation measures of concordance and uncertainty defined above were the outcome measures and I used mixed models to fit each factor individually. Multivariate analyses were not carried out as a number of the factors would be inter-related and we were interested in exploring the influence of each factor on quality, rather than only retaining the strongest characteristics in a multivariate model. Categorical variables were fitted using dummy variables. The dummy variables were a series of dichotomous variables (taking the value 0 or 1). When modelling, I chose one of the categories as the reference value and all of the other dichotomous variables were fitted in the model to compare the effect on the outcome variable compared with the reference category.

The list of factors (Table 14) were discussed and finalised by the project team. The factors were not always applicable (or defined in the same way) for both cross-sectional and longitudinal contrasts, therefore separate models were conducted for the two types of contrast. Some were factors investigated as they are known to affect study quality (as defined for standard meta-analyses), e.g. study design. Some factors were derived during discussions in the project team (anchor, disease, time of cross-sectional contrast and timing of second assessment for longitudinal contrasts).

A couple of factors arose during the expert interviews (Chapter 6). The experts mentioned that if a contrast contained a heterogeneous group of patients it made the review more difficult. In order to investigate this we decided that patients with different

stages of disease best represented the issue of heterogeneity and used this as a measure of heterogeneity for the cross-sectional contrasts.

The experts also raised the issue of anchors not clearly relating to QOL making the reviews hard (Section 6.1.3.3). Although we tried not to include any anchors we felt were unrelated to QOL, the anchors we did include contained a mixture of anchors well known to influence QOL and some lesser known anchors, which although we felt would influence QOL they may not have been well known to all of the experts. Strength of anchor (or familiarity of the anchor(15)) was therefore used as a possible factor affecting the expert review quality. This was based on the team's expert opinion through experience of trials and quality of life research. The categories were defined by Prof Velikova and confirmed by the rest of the project team. This was aimed at identifying anchors that we thought were strongly related to QOL compared to those that although may be related could be largely unfamiliar to the experts.

Dropout or attrition as a potential factor was the subject of much discussion. For standard meta-analyses one may consider studies with a high degree of attrition to be biased. However, here we were looking at cancer studies and some were very long term studies therefore attrition, and high levels of it, would be inevitable. Plus, the interpretation guidelines would ultimately be used for studies of the same nature, i.e. with possibly high levels of attrition. We therefore did not feel these studies should be excluded but it is possible that contrasts with high attrition may have been hard for the experts to judge (since it would require considering the types of patients remaining in the study at that point in time). However, attrition levels for the whole study may be completely irrelevant for some contrasts (e.g. contrasts at baseline or early on in the study) so a different approach to including attrition in the modelling was used. For the cross-sectional contrasts I used the timing of the contrast as an indication of whether this was influencing the concordance or uncertainty (broadly assuming that the later contrasts are more likely to suffer with attrition). For the longitudinal contrasts the definition was easier. I used the timing of the second assessment as one indicator (with similar reasoning as for the cross-sectional factor) but also investigated a more specific measure using the proportion of patients at time 2 compared with that at time 1 in the contrast.

Table 14 Factors possibly affecting expert review quality

Applicable to which contrasts	Factor	Levels	Description	Key for plots
Both	Design	6	Cohort	Coh
			Multiple	Mul
			Non-randomised phase I	NRI
			Non-randomised phase II	NRII
			RCT Phase II	RCTII
			RCT Phase III/Not specified	RCTIII
Both	Disease	13	Brain	Bra
			Breast	Bre
			Colorectal	Col
			Gastro-Intestinal	GI
			Gynaecological	Gy
			Haematological	Hae
			Head and neck	H&N
			Lung	Lun
			Mixed	Mix
			Prostate	Pro
			Testicular	Tes
			Urology/Kidney	Uro
Cross-sectional	Anchor	7	Disease related	Disease
			Treatment/Intervention/ Assessment related	Tmt
			Patient characteristics related	Pat
			Physical function related	Phys
			Symptom/Emotional related	Symp
			Time related	Time
			Survival related	Surv
Cross-sectional	Strength of anchor	3	Well understood	Known
			Unfamiliar	Unknown
			Variable (familiar to some)	Variable

Applicable to which contrasts	Factor	Levels	Description	Key for plots
Cross-sectional	Time (timing of the cross-sectional contrast)	3	Baseline Post-baseline Not specified	Baseline Other NS
Cross-sectional	Disease stage	3	Early Late Mixed	Early Late Mixed
Longitudinal	Timing of second assess-ment	Continuous	Time in months if known	
Longitudinal	Dropout	Continuous	Proportion of patients dropping out from Time 1 to Time 2 for each contrast. Note this is zero if the subset of patients with both times is presented. NB this does not necessarily represent attrition across the study more generally.	

4.4.4 Definition of analysis dataset

The quality assessment described in Section 4.4.3 was used to make decisions on which contrasts represented poor quality evidence and should be excluded from the analysis. The term "full dataset" was used to describe the set of papers/contrasts for which we obtained at least two expert reviews. The "analysis dataset" refers to the subset of these which were deemed good quality evidence and subsequently used in the derivation of the evidence-based guidelines. The analysis dataset consisted of contrasts with:-

- at least two expert reviews (or baseline contrasts between randomised groups)
  and;
- for small, medium and large differences; average expert opinion and the actual QOL difference in the same direction or
- for trivial differences; agreement between reviewers, i.e. weighted averages for each reviewer categorised into zero size class.

4.4.5 Patient review methods

Opinions from patients were sought to develop a method for eliciting informed opinions on QOL differences from patients and to seek to validate the use of experts on our review panel. The full methods and results for this pilot sub-study can be found in Chapter 8.

4.4.6 Sensitivity analyses

Sensitivity analyses were used to test the robustness of the meta-analysis results and subsequent guidelines to variations in the methodology used. The analyses included sensitivity around the imputation methods, the exclusion of poor quality contrasts and the method used for obtaining the random effects variance.

4.5 Evidence-based interpretation guidelines

The meta-analyses were used to estimate an average effect within each size class. Guidelines were then determined using the midpoint between these estimates for each size class. For example, if the estimate for a small effect was 5 points and the estimate for a medium effect was 9 points then the guidelines would recommend a minimum medium effect of 7 points for sample size calculations and interpretation.

4.6 Summary tables and display of results

Chapter 5 contains a summary of the literature search results and details of the studies meeting the inclusion criteria. A flow diagram is used to record the flow of identified articles through the study, including details of reasons for rejection or exclusion at any stage. Summary tables are used to show the characteristics of articles meeting the inclusion criteria with respect to study design, country where carried out,

research question, types of cancer studied as well as a summary of their QOL results overall. Observed mean differences in actual reported QOL scores are summarised to show the range of differences found in the literature. The proportion of contrasts are reported which meet the most commonly used criteria of 10 points for a moderate difference. The subset of randomised treatment comparisons are summarised as these may be informative when considering sample size calculations for RCTs. The Chapter also contains details of the contrasts included in the analysis dataset and a summary of the expert review panel.

Chapter 6 contains the results from the quality assessment including the final definition of the analysis dataset.

The main results (i.e. estimates of trivial, small, medium and large effects from the random effects models) are summarised in Chapter 7. Estimates are displayed as the weighted mean differences (and weighted effect sizes) with 95% confidence intervals as calculated by the random effects models. A summary table and box and whisker plots are used to show the estimates and 95% confidence intervals for trivial, small, medium and large for each subscale.

4.7 Improvements to the original methodology

The methodology for this project built on the EBES project by King *et al(1)*. They used published mean scores from the FACT-G questionnaire and, using meta-analysis techniques, combined these with expert opinion on the size of differences. A number of key improvements were made to the methodology for the purposes of this project.

4.7.1 Exclusion of papers

The EBES project excluded papers with high levels of attrition (>20%) in case of bias. However, since cancer studies, particularly those with long term follow up or for certain types of cancer, suffer from high levels of attrition due to death I felt it may be important to be able to include these studies. They may still be high quality studies and we wanted to include all stages of cancer and as many cancer sites as possible. Attrition bias arises where there is differential attrition across contrast groups (e.g. a higher rate of attrition for one treatment group compared to another) rather than studies with high attrition per se. Since the meta-analysis involves many different contrasts (i.e. not just comparisons of different treatment groups) I felt it was likely that attrition would

not be as much of a concern as it may be for standard applications of meta-analysis. Therefore, instead of excluding studies with high attrition I aimed to include all studies and then investigate attrition as a possible source of heterogeneity.

4.7.2 Expert panel

The EBES project used only three medical oncology experts, who were responsible for the review of papers from a wide variety of cancers and treatments. This had the advantage of the expert panel working very closely with the research team and a consensus process where the individual expert reviews were in clear disagreement was possible. The experts also saw each paper and therefore gained a lot of experience in carrying out the reviews.

I felt that it could improve the quality of the expert reviews if we targeted the review of papers according to the specific area of expertise required for that paper. Therefore the same experts would not review each paper. The panel would include a wider range of expertise such as nurses, radiographers and psychologists rather than just clinicians. This required a much larger panel of experts in order to cover the wide range of cancers and treatments studied in the papers. While a smaller panel has the advantage of the experts being very well trained in the methodology and a consensus process easier to manage there are other advantages to our approach. I hoped that by using a wider range of experts the reviews could be based on an in-depth knowledge in that field and therefore lead to good quality opinions on the appropriate size class for the contrasts.

4.7.3 Quality of the expert reviews

I have incorporated in-depth investigations into the quality of the expert reviews from this project. Using both qualitative and quantitative analysis methods I aimed to explore the process being carried out by the experts to find out what affects the quality of the reviews (and indeed what defines quality in the process). The EBES project used the distance between expert reviews as the lone measure of quality.

4.7.4 Review scale

In the EBES project experts had to judge each contrast by deciding if it would lead to either a trivial, small, medium or large difference in the relevant QOL subscale. In reality, it is likely that when judging a group of patients even an expert would have some uncertainty as to the likely size of the effect. I therefore amended the scale

experts used for their reviews. I allowed the experts to attach a certainty to the judgments. Experts could either decide that the contrast would definitely lead to differences of a particular size or they could spread their expectation across a few of the size classes.

4.7.5 Meta-analysis

The EBES project used basic fixed effect models to combine the contrasts in a meta-analysis. Since there were a wide variety of contrasts being combined (from different cancer sites and based on different clinical anchors) I investigated whether there was heterogeneity before combining the contrasts and then investigated possible sources of the heterogeneity. This is more akin to the method used in standard meta-analyses of studies comparing randomised treatment groups. If heterogeneity is present then the fixed effects models are inappropriate and the random effects model is the best approach to combining the contrasts.

The EBES project had fairly small numbers of contrasts for the longitudinal analyses and therefore any improvements over time were analysed together with deteriorations in time. There is however evidence(86) that relevant differences may vary depending on whether the score is improving or deteriorating. Therefore I felt it was important to try and estimate the size of effect for improvements and deteriorations separately if numbers allowed.

4.7.6 Patient opinions

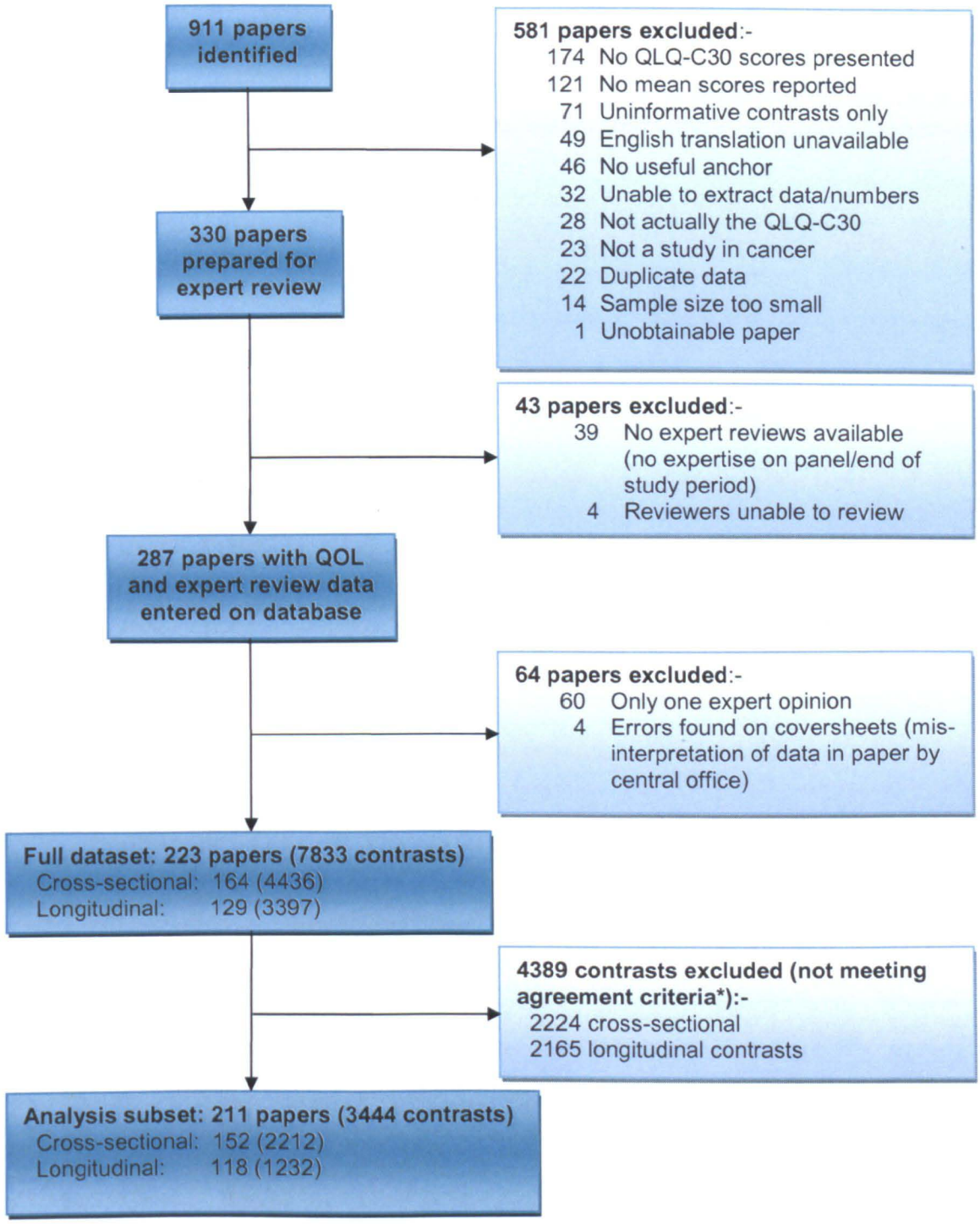
The EBES project did not use the opinions of patients in the development of the interpretation guidelines. This project addresses, for the first time, how patients could contribute to the development of interpretation guidelines. The methodology to obtain patient opinions was developed and piloted as part of this project.

5 Data Extraction and Summaries

5.1 Study flow diagram

The literature search was updated until December 2006. The flow diagram (Figure 8) shows that of an initial 911 papers identified, 330 met the inclusion criteria. The main reasons for rejecting papers were that no scores from the EORTC QLQ-C30 were reported (30%) or, if they were reported, the means of the scores were not available (21%). 287 papers were reviewed by at least one expert or were not sent for review as they only contained randomised contrasts at baseline. The expert review process was conducted from June 2006 to July 2008.

Figure 8 Flow diagram accounting for papers through the project



* Agreement criteria: 1) For trivial contrasts; all experts in agreement. 2) For small, medium or large contrasts; direction of expert opinions in agreement with the observed direction in the paper (e.g. experts and paper both indicate Group A better than Group B).

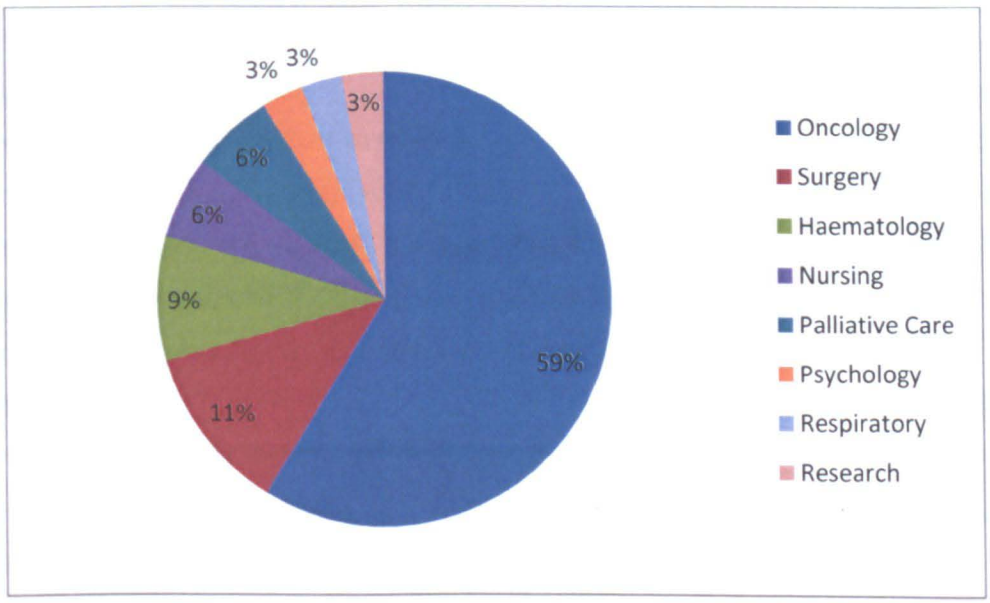
5.2 Expert reviewers

There were 34 expert reviewers involved over the lifetime of the project who reviewed at least one paper. The review panel consisted of oncologists, nurses, surgeons and psychologists. Table 15 shows a summary of the number of papers reviewed by each person (ranging from 1 to 98). Figure 9 shows that less than 60% of the panel were oncologists but the panel also contained surgeons, haematologists, nurses, palliative care specialists, a psychologist, a respiratory physician and a researcher. There were 87 unique combinations of the reviewers (data not shown). The most papers reviewed by any one set of the two/three reviewers was 16.

Table 15 Number of papers reviewed by each reviewer

	Number of papers reviewed
Mean	18.8
Median	12.0
Standard Deviation	20.4
Standard Error	3.5
Max	98.0
Min	1.0
N	34

Figure 9 Expert panel specialities



5.3 Characteristics of eligible papers

Table 16 summarises characteristics of the 330 papers meeting the search criteria. The majority of the 330 papers were cohort or descriptive studies 236 (72%). Randomised controlled trials accounted for another 69 (21%) of the included studies. The remaining papers were reports from multiple studies (17 (5%)) and non-randomised phase I/II studies (8(2%)). The majority of the studies were European (227 (69%)). The research questions addressed by the papers were usually to describe the effects of disease or treatment on QOL (267(81%)), with other research questions such as validation of QOL questionnaires and long term follow up of survivors accounting for less than 5% of the papers each. A wide variety of cancer sites were included in the 330 papers, papers in breast (46(14%)), head and neck (46(14%)) and lung (37(11%)) were the most common disease areas, other areas account for less than 10% of the sample each. In addition to these papers looking at a single cancer site there were 78 (24%) of the papers reporting on multiple cancer sites.

Patients from the 330 studies ranged from age 31-84, with a median age of 61 years. 26 (8%) of studies contained males only, 58 (18%) contained women only and 225 (68%) contained both males and females (these data were missing from the remaining studies).

Table 16 Summary of paper characteristics

	N	%
RESEARCH QUESTION		
Describe effect of disease &/or treatment on HRQOL	267	80.9
Long-term FU of survivors	14	4.2
Develop &/or validate QLQ core module	12	3.6
Comparison of HRQOL instruments	10	3.0
Develop &/or validate QLQ core & disease specific module	6	1.8
Relationship between HRQOL & other variables (e.g. age, sex, survival)	5	1.5
Psychosocial interventions (e.g. nursing, education, counselling programmes)	5	1.5
Develop &/or validate another HRQOL questionnaire	4	1.2
Cross-cultural validation of the QLQ-C30	3	0.9
Reference data for QLQ-C30	1	0.3
Clinical guidelines (eg. developing clinical guidelines)	1	0.3
Cultural issues in HRQOL & health care	1	0.3

	N	%
Economic evaluation	1	0.3
STUDY DESIGN		
Cohort/descriptive study	236	71.5
RCT Phase not specified	46	13.9
Multiple studies	17	5.2
RCT Phase III	17	5.2
RCT Phase II	6	1.8
Non-randomised Phase II	6	1.8
Non-randomised Phase I	2	0.6
REGION		
Europe	227	68.8
Rest of World	56	17.0
USA/Canada	47	14.2
CANCER SITES		
Multiple	78	23.6
Head and neck	46	13.9
Breast	46	13.9
Lung	37	11.2
Haematological	28	8.5
Colorectal	26	7.9
GI	24	7.3
Prostate	22	6.7
Urology/Kidney	7	2.1
Gynaecological	7	2.1
Brain	4	1.2
Testicular	3	0.9
Malignant melanoma	2	0.6
Total	**330**	**100.0**

From the 330 eligible papers, 10682 contrasts were selected for expert review. 57% of the contrasts were cross-sectional and 43% longitudinal.

5.4 Full dataset

The full dataset included any of the eligible papers that had at least two expert reviews by the end of the study period. Of the 330 papers meeting the inclusion criteria 223 remained in the full dataset after the review process. 60 (18%) papers were excluded as they only had one reviewer and 39 (12%) excluded as they had no reviews. Eight further papers had to be excluded during the review process for other reasons. Four papers (1%) were rejected by the reviewers and four (1%) had errors on the coversheets sent to reviewers (Figure 8). The papers rejected by reviewers were due to the wide range of patients in the sample which reviewers felt made it impossible to judge what the QOL differences might be.

The full dataset had very similar paper and patient characteristics to the 330 eligible papers described above (data not shown). There were 7833 contrasts (57% cross-sectional, 43% longitudinal).

5.5 Analysis dataset

The analysis dataset is defined in Section 4.4.4. Once the criteria for good quality contrasts had been applied there were 152 papers contributing 2212 cross-sectional contrasts and 118 papers contributing a total of 1232 longitudinal contrasts for analysis (Figure 8). A total of 7238 contrasts were excluded (68% of the 10682 original eligible contrasts). Table 17 shows a breakdown of the reasons for excluding contrasts from the analysis dataset. Exclusion criterion 1 refers to the trivial difference size class, where a contrast was excluded if reviewers did not all rate the contrasts as trivial. Exclusion criterion 2 refers to the exclusion of contrasts from the small, medium and large size classes (where there was disagreement between the direction of the expert overall opinion and the actual QOL scores).

Table 17 Reasons for exclusions from analysis subset

Number of contrasts excluded (% of those in eligible dataset)	Reason for exclusion			Total excluded	**Remaining in analysis dataset**
	Insufficient number of expert reviews	Exclusion criterion 1	Exclusion criterion 2		
Cross-sectional (from 6089 eligible)	1653 (27%)	1436 (24%)	788 (13%)	3877 (64%)	**2212 (36%)**
Longitudinal (from 4593 eligible)	1196 (26%)	1557 (34%)	608 (13%)	3361 (73%)	**1232 (27%)**
Total excluded (from 10682 eligible)	2849 (27%)	2993 (28%)	1396 (13%)	7238 (68%)	**3444 (32%)**

Table 18 summarises characteristics of papers in the analysis dataset. The characteristics were very similar to those of the 330 eligible papers described in Section 5.3. The majority of the 211 papers were cohort or descriptive studies 149 (71%). Randomised controlled trials accounted for another 51 (24%) of the included studies. The remaining papers were reports from multiple studies (7 (3%)) and non-randomised phase I/II studies (4 (2%)). The majority of the studies were European (150 (71%)). The research questions addressed by the papers were usually to describe the effects of disease or treatment on QOL (175 (83%)), with other research questions such as validation of QOL questionnaires and long term follow up of survivors accounting for a maximum of 5% of the papers each. A wide variety of cancer sites are included in the analysis dataset, papers in breast (42 (20%)), lung (35 (17%)), head and neck (28 (13%)) and colorectal (22 (10%)) were the most common disease areas. Other areas accounted for less than 10% of the sample each. In addition to these papers looking at a single cancer site there were 24 (11%) of the papers reporting on multiple cancer sites.

Patients from the 211 studies ranged from age 31-74, with a median age of 62 years. 21 (10%) of studies contained males only, 45 (22%) contained women only and 136 (67%) contained both males and females (these data were missing from the remaining studies).

Table 18 Summary of paper characteristics (analysis dataset)

	N	%
RESEARCH QUESTION		
Describe effect of disease &/or treatment on HRQOL	175	82.9
Long-term FU of survivors	11	5.2
Develop &/or validate QLQ core module	8	3.8
Comparison of HRQOL instruments	5	2.4
Psychosocial interventions (e.g. nursing, education, counselling programmes)	4	1.9
Develop &/or validate QLQ core & disease specific module	3	1.4
Develop &/or validate another HRQOL questionnaire	2	0.9
Cross-cultural validation of the QLQ-C30	2	0.9
Cultural issues in HRQOL & health care	1	0.5
STUDY DESIGN		
Cohort/descriptive study	149	70.6
RCT Phase not specified	30	14.2
RCT Phase III	15	7.1
Multiple studies	7	3.3
RCT Phase II	5	2.4
Non-randomised Phase II	3	1.4
RCT Phase III crossover	1	0.5
Non-randomised Phase I	1	0.5
REGION		
Europe	150	71.1
Rest of World	34	16.1
USA/Canada	27	12.8
CANCER SITES		
Breast	42	19.9
Lung	35	16.6
Head and neck	28	13.3
Multiple	24	11.4

	N	%
Colorectal	22	10.4
Haematological	19	9.0
Prostate	18	8.5
GI	10	4.7
Urology/Kidney	5	2.4
Brain	4	1.9
Testicular	3	1.4
Gynaecological	1	0.5
Total	**211**	**100.0**

Thirty-two percent of the papers had two reviews and 68% had three reviews (Table 19). The number of contrasts ranged from 1 to 24 per paper with a median of 2. The sample size of the included papers ranged from 10 to 2640 patients with a median of 133 patients. The global quality of life scale was the most frequently reported subscale (Table 20). The financial difficulties subscale was often omitted from the results.

Table 19 Number of papers/contrasts with at least two reviewers (analysis dataset)

Number of reviewers	Contrasts	% of all contrasts	Papers	% of all papers
Two	1372	39.8	67	31.8
Three	2072	60.2	144	68.2
Total	3444		211	

Table 20 Number of contrasts by subscale (analysis dataset)

Subscale	Cross-sectional		Longitudinal	
	N	%	N	%
Appetite Loss (AP)	134	6	76	6
Cognitive Functioning (CF)	163	7	66	5
Constipation (CO)	128	6	70	6
Diarrhoea (DI)	135	6	75	6
Dyspneoa (DY)	112	5	58	5
Emotional Functioning (EF)	182	8	116	9

Subscale	Cross-sectional		Longitudinal	
	N	%	N	%
Fatigue (FA)	150	7	102	8
Financial Impact (FI)	82	4	37	3
Nausea and Vomiting (NV)	131	6	80	6
Pain (PA)	148	7	92	7
Physical Functioning (PF)	188	8	91	7
Global QOL (QL)	195	9	118	10
Role Functioning (RF)	171	8	89	7
Social Functioning (SF)	182	8	97	8
Insomnia (SL)	111	5	65	5
All	2212		1232	

I initially coded the contrasts under 50 individual anchors, using descriptions for the anchors from the source papers. I then collated these using seven broad categories for the description of the anchor (Table 21). Time-related and treatment-related anchors were the most common.

Table 21 Frequency of anchors (analysis dataset)

Anchor (categorised)	Examples of individual anchors or contrasts	Number of papers	Number of contrasts
Treatment/Intervention/ Assessment related	Graft versus host disease, Previous chemotherapy, Responsibility for follow up, Treatment (non-rand), Treatment (rand), Type of surgical technique, Use of pain killers, Use of psychosocial support	87	875
Time related	Cross-sectional: <1yr post BMT vs >1yr post BMT, active treatment group versus follow up group Longitudinal: Baseline vs 2nd cycle, Pre vs post treatment, Baseline vs 6mths follow up	123	1275
Disease related	Clinical phase of disease, Co-	54	628

Anchor (categorised)	Examples of individual anchors or contrasts	Number of papers	Number of contrasts
	morbidity, Disease stage, Presence of metastases, Presence of stoma, Primary disease, Response status, Time since diagnosis, Tumour site		
Patient characteristics related	Age, Gender, Marital Status	24	239
Physical function related	Arm problems, Performance status, Erectile dysfunction, Anal function, Urinary continence, Dysphasia, Motor deficit	16	317
Symptom/Emotional related	Anxiety, Confusion, Pain, Weight loss, Insomnia, Neurological status, Symptom reporting	9	92
Survival related	All patients vs completers, Baseline comparisons of survivors at 3 months vs non-survivors at 3 months	2	18

5.6 Data extraction/Imputation

Standard deviations could be directly extracted from the paper or calculated from the information provided for 57% of the cross-sectional contrasts and 50% of the longitudinal contrasts in the analysis dataset (Table 22).

Table 22 Level of imputation required for analysis dataset

	SD available	SD imputed
Cross-sectional	1273 (58%)	939 (42%)
Longitudinal	613 (50%)	619 (50%)

The impact of imputation on the analysis results was investigated later via sensitivity analyses, see Section 7.3.1 for details.

5.7 Summary of QOL data from papers

The QOL data was summarised for the analysis dataset in order to show the range of QLQ-C30 scores from the contrasts in the final analysis. Mean differences for cross-sectional comparisons (Table 23) ranged from 0 to 64 points with a median of 5.8 points. Across subscales the medians ranged from 3.0 (DI and NV subscale) to 9.7 (RF subscale). For longitudinal comparisons (Table 24) the difference score ranged from -54 (a deterioration in the DI subscale) to 50 points (an improvement in the PA subscale), with a median of 1. Across subscales the medians range from -4 (PF subscales) to 6 (EF subscale). Overall 30% and 27% of the contrasts had QOL differences of more than 10 points for cross-sectional and longitudinal contrasts respectively (Figure 10 and Figure 11).

The RF subscale had the highest proportion of QOL scores of 10 or more for both cross-sectional and longitudinal contrasts. There were some subscales with scores which deteriorated over time on average (FA, NV, PF and RF) and some which improved over time on average (EF, PA and SL). EF in particular had a higher average QOL difference when compared with the other subscales, with an average mean difference of 6 points (effect size 0.27) compared to the average over all subscales of 1 (effect size 0.04).

Baseline scores for the longitudinal comparisons are summarised in Table 25. This shows that some subscales had an average score at baseline which represents a particularly low level of symptoms (NV, DI and CO) or a high level of functioning (CF, PF and SF) which actually leaves little room for improvements over time.

Table 23 Analysis dataset: Cross-sectional mean differences and effect sizes

All subscales	Mean difference						Effect size						
	Mean	Median	Standard Deviation	Standard Error	Maximum	Minimum	Mean	Median	Standard Deviation	Standard Error	Maximum	Minimum	N
AP	8.8	5.9	8.6	0.8	37.7	0.0	0.36	0.25	0.36	0.03	1.86	0.00	134
CF	7.0	5.0	6.5	0.5	33.1	0.0	0.33	0.24	0.36	0.03	2.93	0.00	163
CO	7.8	4.4	8.5	0.8	45.7	0.0	0.31	0.21	0.33	0.03	1.76	0.00	128
DI	4.2	3.0	4.2	0.4	22.0	0.0	0.20	0.15	0.19	0.02	0.90	0.00	135
DY	7.1	5.0	6.4	0.6	31.6	0.0	0.26	0.19	0.22	0.02	1.18	0.00	112
EF	7.2	5.0	6.9	0.5	42.1	0.2	0.31	0.22	0.31	0.02	1.90	0.01	182
FA	10.5	7.8	8.9	0.7	41.0	0.0	0.42	0.32	0.38	0.03	2.35	0.00	150
FI	5.7	4.3	5.1	0.6	26.3	0.0	0.22	0.18	0.19	0.02	0.90	0.00	82
NV	5.7	3.0	6.8	0.6	38.0	0.0	0.31	0.20	0.34	0.03	2.10	0.00	131
PA	9.7	7.0	9.5	0.8	49.0	0.0	0.37	0.27	0.35	0.03	1.75	0.00	148
PF	11.6	8.3	11.2	0.8	60.0	0.0	0.51	0.36	0.55	0.04	4.21	0.00	188
QL	8.5	5.5	7.7	0.6	35.0	0.0	0.37	0.25	0.33	0.02	1.67	0.00	195
RF	13.4	9.7	12.7	1.0	64.0	0.1	0.48	0.29	0.71	0.05	5.92	0.00	171
SF	9.2	7.0	8.1	0.6	50.0	0.0	0.34	0.29	0.30	0.02	1.83	0.00	182
SL	6.8	5.0	6.2	0.6	29.3	0.2	0.25	0.17	0.31	0.03	2.13	0.01	111
All	8.5	5.8	8.7	0.2	64.0	0.0	0.35	0.24	0.39	0.01	5.92	0.00	2212

Table 24 Analysis dataset: Longitudinal contrasts mean differences and effect sizes

All subscales	Mean difference						Effect size						
	Mean	Median	Standard Deviation	Standard Error	Maximum	Minimum	Mean	Median	Standard Deviation	Standard Error	Maximum	Minimum	N
AP	-0.8	1.0	13.1	1.5	30.6	-41.0	-0.02	0.03	0.47	0.05	1.41	-1.34	76
CF	0.0	-1.0	6.7	0.8	20.8	-17.5	0.02	-0.05	0.39	0.05	1.69	-0.76	66
CO	-0.2	1.6	9.0	1.1	19.0	-30.3	-0.01	0.06	0.33	0.04	0.68	-1.12	70
DI	-1.5	-0.1	9.1	1.1	27.0	-54.0	-0.08	-0.00	0.42	0.05	1.25	-2.49	75
DY	-0.3	1.0	8.6	1.1	19.0	-26.7	-0.00	0.04	0.32	0.04	0.69	-0.86	58
EF	5.8	6.0	7.9	0.7	22.4	-29.0	0.27	0.30	0.39	0.04	1.89	-1.45	116
FA	-3.8	-3.8	9.3	0.9	22.0	-27.4	-0.14	-0.15	0.41	0.04	1.45	-1.50	102
FI	-0.5	0.8	5.8	1.0	12.7	-14.9	-0.03	0.03	0.24	0.04	0.45	-0.58	37
NV	-4.1	-3.0	10.3	1.2	24.3	-34.0	-0.19	-0.19	0.50	0.06	0.94	-1.56	80
PA	2.9	4.0	10.6	1.1	50.3	-22.4	0.13	0.15	0.45	0.05	2.61	-1.15	92
PF	-4.1	-4.0	9.7	1.0	13.8	-45.0	-0.17	-0.15	0.41	0.04	0.57	-1.80	91
QL	0.1	2.8	9.6	0.9	22.0	-36.3	0.01	0.12	0.43	0.04	1.17	-1.53	118
RF	-3.2	-3.0	13.5	1.4	20.1	-50.0	-0.11	-0.09	0.45	0.05	0.64	-1.50	89
SF	-1.4	-1.0	9.5	1.0	16.1	-33.4	-0.05	-0.03	0.36	0.04	0.59	-1.18	97
SL	3.7	4.0	8.6	1.1	17.5	-28.2	0.13	0.13	0.30	0.04	0.63	-1.03	65
All	-0.4	1.0	10.2	0.3	50.3	-54.0	-0.01	0.04	0.42	0.01	2.61	-2.49	1232

Figure 10 Analysis dataset: Cross-sectional contrasts – proportion with QOL differences of 10 or more points

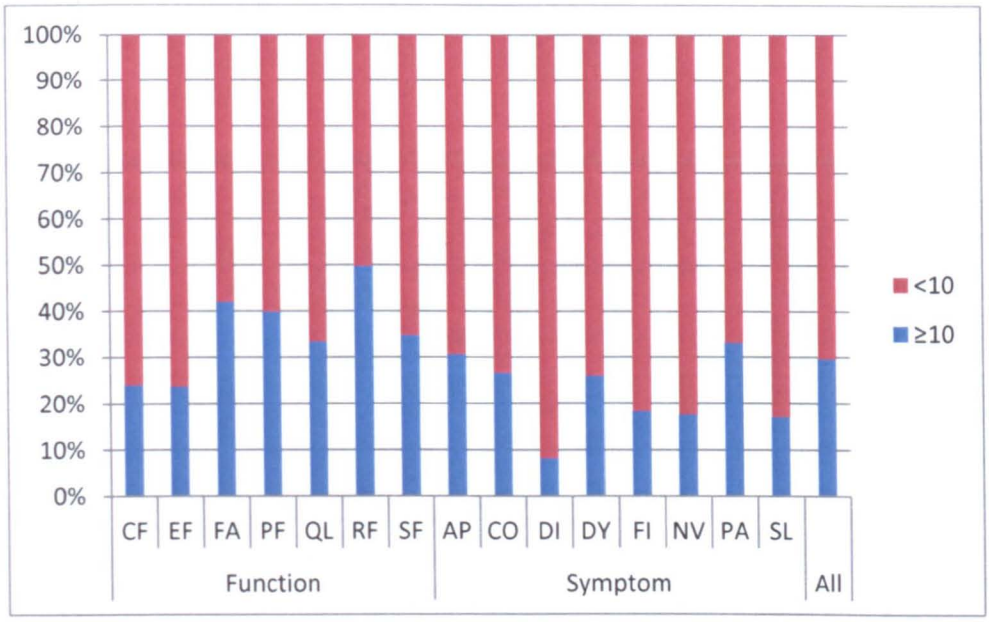


Figure 11 Analysis dataset: Longitudinal contrasts – proportion with QOL differences of 10 or more points

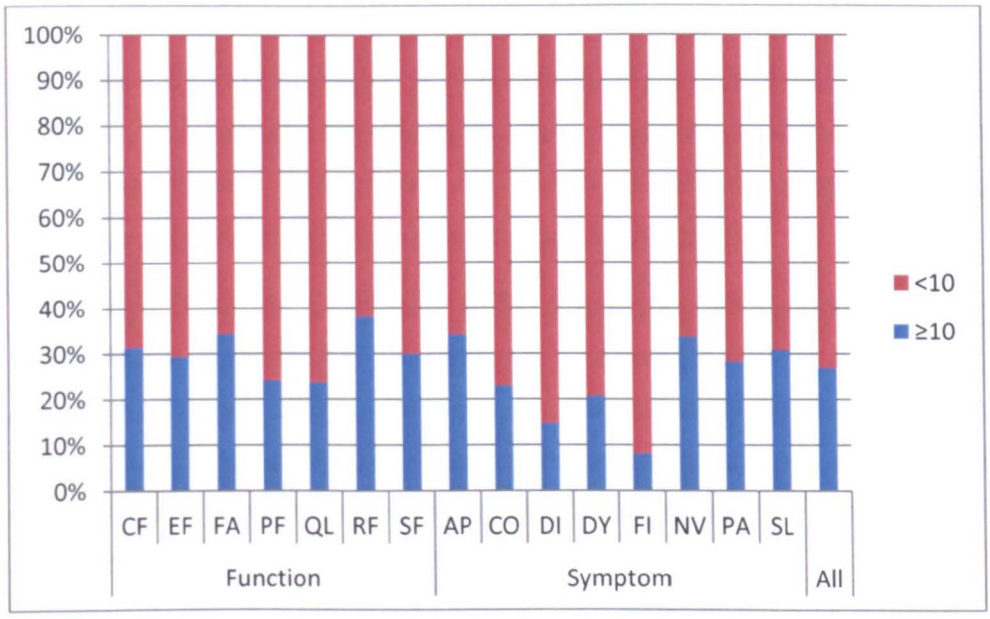


Table 25 Analysis dataset: Baseline scores for the longitudinal contrasts

Mean baseline scores for each contrast		Mean (overall)	Median	SD	SE	Max	Min	Number of contrasts with mean baseline score provided
Function	CF	81.2	83.5	10.3	1.4	96.7	42.7	57
	EF	69.2	70.0	9.0	1.0	89.9	44.0	88
	PF	77.8	78.9	11.0	1.3	96.3	46.0	75
	QL	62.8	63.0	8.2	0.8	83.1	40.0	99
	RF	71.3	73.4	13.2	1.6	95.7	34.0	70
	SF	77.1	78.2	9.1	1.0	94.7	51.0	76
Symptom	AP	22.7	17.0	16.7	2.1	87.6	3.2	61
	CO	19.8	15.5	15.4	2.0	80.5	0.0	57
	DI	13.8	8.7	17.2	2.2	91.0	0.4	62
	DY	24.6	17.3	18.3	2.7	72.0	7.0	46
	FA	34.6	33.0	12.5	1.4	81.0	6.5	83
	FI	18.4	13.9	14.1	2.7	53.3	1.2	27
	NV	9.0	7.0	7.4	0.9	36.0	2.0	67
	PA	27.3	25.4	14.4	1.7	88.1	5.0	73
	SL	29.2	28.0	6.8	0.9	46.0	12.0	55

5.7.1 Randomised treatment contrasts

The subset of contrasts from randomised controlled trials in the analysis dataset were also summarised separately from the other cross-sectional contrasts. The mean differences arising from these randomised treatment comparisons could be very informative when planning sample sizes for new studies.

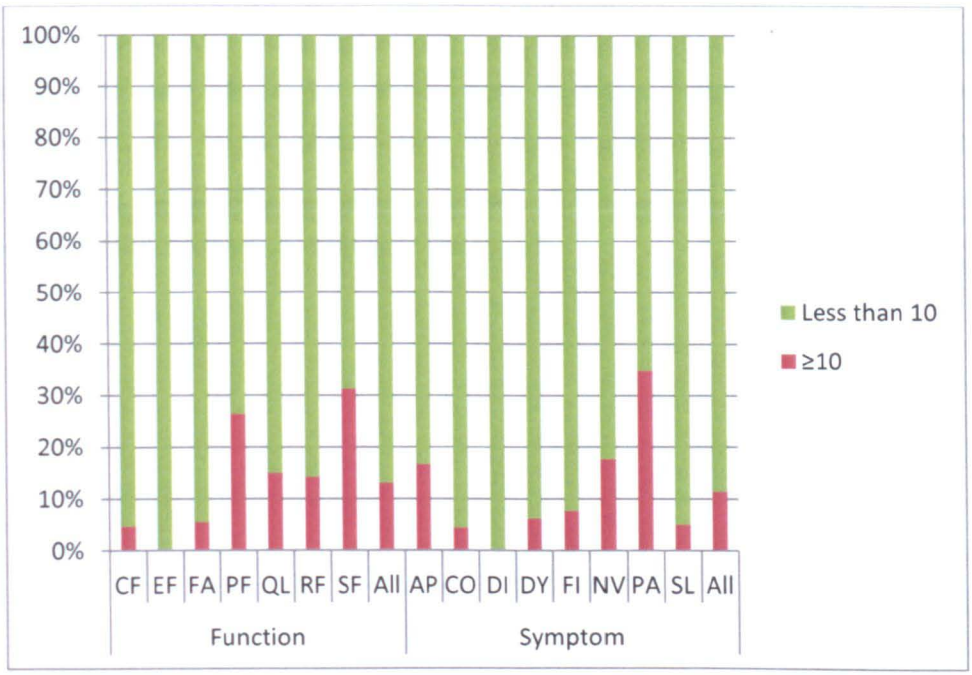
Table 26 shows a summary of the differences arising from randomised treatment comparisons (excluding any randomised treatment contrasts at baseline). There were 297 contrasts from 29 papers which compared two randomised treatments. Note that the number of contrasts was high compared to the number of papers as treatments could be compared at multiple points in time and more than two treatments could be in the randomisation.

Figure 12 shows the proportion of contrasts for each subscale with the larger improvements (10 or more points difference between the treatment groups). Overall there were 12% of the RCT treatment contrasts achieving 10 or more points difference. The PA and SF subscales were the most likely to show the larger differences.

Table 26 Analysis dataset: Randomised treatment comparisons

All subscales	Mean difference							Effect size					
	N	Mean	Median	Standard Deviation	Standard Error	Maximum	Minimum	Mean	Median	Standard Deviation	Standard Error	Maximum	Minimum
AP	134	6.6	4.2	10.5	0.9	37.7	-14.0	0.28	0.20	0.43	0.04	1.86	-0.41
CF	163	4.4	3.1	8.6	0.7	33.1	-25.1	0.22	0.17	0.43	0.03	2.93	-0.59
CO	128	5.4	3.0	10.2	0.9	45.7	-17.0	0.20	0.12	0.40	0.04	1.76	-0.90
DI	135	2.3	1.5	5.5	0.5	22.0	-12.0	0.11	0.09	0.26	0.02	0.90	-0.76
DY	112	4.5	4.0	8.5	0.8	31.6	-20.0	0.15	0.15	0.31	0.03	1.18	-0.75
EF	182	5.3	4.0	8.4	0.6	42.1	-22.8	0.23	0.18	0.37	0.03	1.90	-1.04
FA	150	9.2	7.4	10.2	0.8	41.0	-21.9	0.37	0.30	0.43	0.03	2.35	-0.81
FI	82	2.4	1.6	7.2	0.8	26.3	-14.1	0.09	0.07	0.28	0.03	0.90	-0.71
NV	131	4.3	2.9	7.8	0.7	38.0	-18.8	0.23	0.17	0.40	0.03	2.10	-0.63
PA	148	8.5	6.8	10.6	0.9	49.0	-21.0	0.32	0.25	0.39	0.03	1.75	-0.75
PF	188	10.9	8.0	11.9	0.9	60.0	-11.8	0.48	0.34	0.58	0.04	4.21	-0.55
QL	195	7.8	5.2	8.4	0.6	35.0	-11.9	0.34	0.23	0.37	0.03	1.67	-0.66
RF	171	12.0	9.0	14.1	1.1	64.0	-19.8	0.44	0.27	0.74	0.06	5.92	-0.67
SF	182	8.0	6.9	9.3	0.7	50.0	-15.6	0.30	0.26	0.35	0.03	1.83	-0.70
SL	111	4.8	4.0	7.8	0.7	29.3	-13.0	0.18	0.12	0.35	0.03	2.13	-0.44
All	2212	6.8	5.0	10.0	0.2	64.0	-25.1	0.28	0.20	0.45	0.01	5.92	-1.04

Figure 12 Analysis dataset: Randomised treatment comparisons – proportion with QOL differences of 10 or more points



5.8 Summary and conclusions

Not all papers identified in the literature search were appropriate for the project. The majority of papers which were excluded were due to the exclusion criteria we defined *a priori*. A smaller proportion of the papers were subsequently excluded during or after the expert review process. Although we had a large expert panel with a wide range of expertise, we could not identify appropriate experts for some of the papers.

Following the review process, more than half of the reviewed contrasts violated the criteria for the analysis subset and were subsequently excluded from the analysis. The criteria for inclusion in the analysis subset were defined post hoc which is not ideal (full details and discussion follow in Chapter 6) but they were derived with the aim of only including the best quality contrasts in the analysis. Despite the large proportion of exclusions the papers included in the final analysis had very similar characteristics (e.g. study design, research question, cancer sites and so on) to the complete set of eligible papers.

In common with other meta-analyses, variance data could not be extracted for all of the contributing contrasts. Standard methods were employed to impute these data and the possible impact of imputation was later investigated using sensitivity analyses.

Around 30% of the contrasts in our analysis met Osoba's definition of a moderate difference (10 or more points). When the contrasts were limited to just the treatment comparisons from RCTs this proportion was considerably lower, with only 12% of the contrasts achieving differences of this size. Some subscales were more likely than others to yield the larger differences (eg role functioning, physical functioning and fatigue when comparing between groups; pain, social functioning and physical functioning when assessing change over time).

6 Quality assessment

Chapter 6 is divided into two parts; part 1 reporting the qualitative interviews with the experts and part 2 reporting the quantitative analysis used to define the contrasts for the analysis dataset.

6.1 Part 1: Expert interviews

6.1.1 Aims

The aims of the interviews were:

a) To find out how reviewers had approached the task

b) To explore any possible issues with reviewing certain papers or subscales

c) To generate possible factors affecting the quality of the judgements and use them in subsequent quantitative assessment of quality.

6.1.2 Methods

6.1.2.1 Interview

A telephone interview with each expert on the panel was planned. The interview consisted of 12 questions (Table 27).

Table 27 Expert interview questions

1. Before starting this project did you already have an idea of what difference in QLQ-C30 scores would be clinically relevant? If yes:- a) What is it and what is this based on? b) Is it the same for all subscales?
2. Were there particular study designs or types of papers that you found hard to make a judgement on or that you were less sure of your answers on?
3. Were there particular clinical anchors that you found hard to make a judgement on or that you were less sure of your answers on?
4. Did you find it easier to think about the cross-sectional comparisons or those over time? Why?

5. Were there particular subscales of the QLQ-C30 that were hard to make judgements on or you were less sure of your answers on?
6. Can you describe your approach to the task? i.e. what process did you go through with each paper?
7. Which aspects of the paper that you could see did you use most to help you fill in the tables?
8. How long would you say on average each paper took you to review?
9. Do you feel that experience of using the QLQ-C30 in cancer patients is enough to predict the size of difference in QOL scores in this way?
10. Do you think there is a minimal level of experience that would make this task possible?
11. Do you think it is possible to use published data at all to inform clinical interpretation of QOL scores? If so do you think there is an alternative way of utilising the published data?
12. Additional comments

The first question was designed to find out if reviewers were familiar with any of the literature on how to interpret QOL scores. We wanted to see if their judgements for this project might be based on work already done in the area of interpretation of EORTC QLQ-C30 scores specifically or QOL scores more generally. We were keen to avoid reviewers trying to guess what the quality of life scores might be in the papers and then making their judgements using a pre-conceived idea of what difference was clinically relevant, for example 10 points. We felt that if reviewers were approaching the task in this way it could compromise the aim of this project which is to look at a new approach to interpretation.

Questions 2-5 were designed to highlight if there were any particular types of study, comparison or subscales which reviewers found hard to make a judgement on. These questions were included in order to identify factors that may indicate poor quality contrasts and to inform future projects of this nature by helping to screen papers for inclusion and highlight subscales that may not be as familiar to the reviewers and therefore not appropriate for this type of review. We felt the answers to these questions could also help to explain disagreements in the reviewers' judgements and aid interpretation of the main analysis.

Reviewers were asked to read a handbook(68) which explained the aim of the reviews and included instructions on how to fill in the coversheets. However it was

anticipated that different reviewers may approach the task in different ways. Two questions (6/7) were designed to explore this and to see if there were specific sections from the papers that reviewers were using.

Questions 8-11 were included to obtain the expert panels views on the project and methodology behind it as well as inform future studies using the same methodology on how the criteria for an expert may be defined.

6.1.2.2 Sample and timing of the interview

Interviews were conducted when reviewers had completed the review of some papers but, where possible, were still involved in the review process. All reviewers were invited to participate but a minimum of two reviewed papers was preferred prior to the interview.

6.1.2.3 Data collection and analysis

For each interview the reviewer's gender, area of expertise, the type and number of papers reviewed for the project were recorded along with date/time and duration of the interview. Telephone interviews were recorded and transcribed verbatim. The interviews were carried out by myself and the research assistant.

I used thematic content analysis (87) to organise the data from each question and group extracts of text. This involved forming a table for each question containing all the answers. I then extracted themes from the text. Results for some questions were also summarised quantitatively.

6.1.3 Results

6.1.3.1 Sample

Thirteen reviewers were interviewed out of the 29 returning at least two papers (45%), representing 38% of the 34 experts involved over the lifetime of the project. Nine were interviewed by telephone as planned and four by returning a questionnaire (as a suitable interview time could not be arranged). Interviews were between 15 and 26 minutes long, with the majority taking 15 minutes. The characteristics of the interviewed reviewers are summarised in Table 28.

Table 28 Interviewee characteristics

Sex (M:F)	54%:46%
Background	
Psychology	2 (15%)
Respiratory medicine	2 (15%)
Oncology	2 (15%)
Neuro-oncology	1 (8%)
Surgery	1 (8%)
Radiotherapy	1 (8%)
Haematology	1 (8%)
Research fellow	1 (8%)
Research coordinator	1 (8%)
Public Health	1 (8%)
Categories of papers reviewed (more than one area may have been reviewed per reviewer)	Number of interviewees (%)
Brain	1 (8%)
Breast	5 (38%)
Colorectal	2 (8%)
Complementary therapy	1 (8%)
Gastro-Intestinal	1 (8%)
Gynaecology	2 (15%)
Haematology	1 (8%)
Lung	5 (38%)
Pain management	1 (8%)
Palliative Care	2 (15%)
Prostate	1 (8%)
Psychosocial oncology	2 (15%)
Testicular	1 (8%)

6.1.3.2 Pre-conceived ideas of clinically relevant differences

9/13 reviewers answered yes, indicating they had some idea of what a clinically relevant difference in scores would be. Four of these reviewers referred to Osoba's work(49), although none gave specific details in terms of the number of points deemed to be clinically relevant. One also mentioned Cohen's work(31). One reviewer used a

10% change as significant based on their previous work (although this was not specifically based on the EORTC QLQ-C30). The remaining four reviewers had their own idea of a clinically relevant difference based on their previous clinical or trials experience.

Four reviewers with an idea of clinical relevance prior to starting the project said this was not the same for all subscales, two reviewers used the same measure of clinical relevance regardless of subscale, one reviewer was unsure if it should be the same for all subscales and two reviewers were non-responders for this question.

6.1.3.3 Feedback on the types of papers and comparisons reviewed

There were seven main themes emerging from question 2, these are summarised in Table 29 along with quotes from the interviews.

Table 29 Study designs or papers hard to make a judgement on

Theme (number of interviews extracted from)	Quotes
Unfamiliar treatments/interventions(3)	I was more confident with studies that had radiotherapy being looked at as that's what I clinically deal with every day, so I am a lot more familiar with radiotherapy. REVIEWER15 Because I am a psychologist... I am not familiar with the intricacies of each type of treatment as for example a medical doctor would be. So I found that quite difficult, and unless for some reason I was actually familiar with that through my work it would just be through the course of the work and not like, for example, a medical oncologist would be. I tended to email back and email that it was out of my expertise...So it had nothing to do with the study design, but more the types of treatments – whether they were medically based treatments that I had no familiarity with. REVIEWER36
Heterogeneous patient populations (2)	Papers that had a homogenous population were easiest... The most difficult were those with populations that included subjects with both localised & metastatic disease. REVIEWER24

Theme (number of interviews extracted from)	**Quotes**
Complex trials, e.g. trials with several stages (2)	I found the questionnaire very useful in randomised controlled trials, I found it much more difficult in trials which had several stages to them for example if they are comparing a run-in period and then a treatment period. REVIEWER38 Not really. Most of the studies I had were all very similar. Most had a primary outcome that were looking at treatment. All relatively simple comparative studies. REVIEWER21
High attrition rate (2)	Another thing which is difficult, and which is a problem with all these questionnaires is when you're looking at very sick people particularly studies that looked at patients receiving best supportive care versus no treatment, or best supportive care versus a single treatment because the attrition rate is so very high. Therefore all these QOL questionnaires select out the better patients because the poorer patients are too unwell to answer. REVIEWER38
Studies where patients could move in and out of groups over time (1)	... for example patients with lymphedema following surgery who some years didn't have this symptom and vice versa. In my mind there is some degree of cancelling out in papers like this. REVIEWER30
Poorly designed studies and small sample sizes (1)	If it's been badly designed in terms of the research question, then that messes up the quality of life scores. REVIEWER12
All hard (2)	All hard, only have history with patients and they don't tell everything to docs REVIEWER05

Table 30 shows the results from question 3. 5/13 reviewers did not consider any anchors harder to judge than others. Two found all were difficult, although one of these only reviewed two papers. Anchors highlighted as difficult were time, side effects, those (in the reviewers opinion) with little or no relationship to QOL and contrasts comparing

groups fairly close together (e.g. next to each other on a scale, such as ECOG 0 versus ECOG 1, rather than comparison of groups at opposite ends of the scale).

Table 30 Difficulty of different anchors

Theme (number of interviews extracted from)	**Quotes**
No difference in difficulty (5)	Not really. The anchors were all relatively clear. None harder than others. REVIEWER21
All difficult (2)	All difficult. Side effects particularly variable and patient degree of well-being after treatment also variable REVIEWER05
Time anchors hard (1)	The time ones were most difficult. REVIEWER30
Side effects (1)	Side effects from drugs, particularly fatigue are very variable and degree of well-being a patient may have after treatment also varies. REVIEWER05
Anchors with no relationship to QOL hard (1)	Easy were age, sex, disease stage, disease type and treatment. Hard ones were silly things like some biological marker that would have no impact on QOL. REVIEWER12
Anchors with closely defined groups hard (1)	...obviously it is easier to make a determination between and ECOG1* and ECOG4* patient, but sometimes the difference between ECOG2* and 3* were more difficult. And ... it was a lot easier thinking about a 40 year old and an 80 year old, but trying to think about 40-50 year old compared to 50-60 year old I found that I was more indecisive. REVIEWER15 **referring to ECOG performance status stages*

Table 31 summarises results from question 4, indicating whether longitudinal or cross-sectional contrasts were more difficult to judge. 3/13 reviewers had no preference, four found the time comparisons easier and four found the cross-sectional comparisons were easier. One answer was ambiguous and one reviewer could not remember.

Table 31 Preference for cross-sectional or longitudinal comparisons

Theme (number of interviews extracted from)	Quotes
Longitudinal easier (4)	Time comparisons were easier than cross-sectional comparisons as we follow up patients over a period of time. REVIEWER05 Over time was easier. Helped to think of the QLQ scales in terms of clinical changes over time. Over time you can estimate what these might be – getting better or getting worse. Harder to think how groups might be different cross-sectionally. REVIEWER21 I thought the ones over time were much easier than the cross sectional but that was maybe related to the study design of the papers, because some of the papers contain very heterogeneous groups, and to do a cross-sectional analysis of very heterogeneous populations I think it's flawed with all sorts of hazard. You're often looking at a mean value of heterogeneous groups and there is a great risk that they don't really relate to the group as a whole. Longitudinal easier, but again, the attrition rate is a huge influence. REVIEWER38 Those over time. I suppose again it comes back to the treatment thing, I had to think very very hard on some papers about the treatment, and so over time you're making a more general assumption on what is happening to a person's Quality of Life and so there are general assumptions that I suppose apply across all treatments, and that's why I found the time thing easier as opposed to comparing two different groups. This relates back to my lack of familiarity of the treatment types (as cross sectional comparisons are based on treatments generally). REVIEWER36

Theme (number of interviews extracted from)	Quotes
Cross-sectional easier (4)	The time ones were difficult. When I went back to check my consistency, I found on the time ones this was where I was making mistakes. Time ones are not intuitive. The cross-sectional ones I knew what the differences should be. Time ones I had to really think what am I comparing with what? REVIEWER30 Cross-sectional easier, as a direct comparison between groups. Longitudinal more difficult, as (i) need to consider QOL effects of both toxicity & relief of cancer symptoms (see #2), and because of adaptation which will reduce sensitivity to detect changes in QOL. REVIEWER24 I think the cross-sectional ones were easier because it was to consider the sample characteristics and clinical scenario at a single time point REVIEWER06 Cross sectional. I was more confident with these than longitudinal. I think this was simply because I am more familiar within a clinical setting. When we're talking to patients, we tend to talk about those differences at one time point rather than talking about how they feel now and how they may feel in 2 years' time. Generally we're talking about comparing one treatment to another or no treatment to another rather than discussing with patients a prediction of how they'll feel over the next 2-3 years, so I am not as familiar with doing that. Most of these studies work prospectively, and the patients' memory and the way they're thinking about the Quality of Life may affect the outcome. REVIEWER15
No preference (3)	Didn't bother me which. REVIEWER12

Table 32 summarises results from question 5, indicating whether any of the subscales were harder to judge. 3/13 reviewers said no. The other reviewers highlighted that generally subscales describing symptoms were found the least difficult. Subscales felt to be related to the disease or treatment were found easier and in particular contrasts which were specific to the disease rather than more general made it easier.

Table 32 Difficulty of different subscales

Theme (number of interviews extracted from)	Sub-theme	Quotes
Some subscales harder (9)	Financial difficulties hard to judge (4)	Yes, financial difficulty, constipation, diarrhoea, dyspnoea. Most patients were covered by health insurance, but cancer treatment may also cause some financial difficulties, so I am not sure how much. REVIEWER37 In general I focused on the things that were clinical or disease related like cough or pain. Things like Financial might be the same and you'd 20% in all of the boxes because you haven't got a clue what might happen... REVIEWER21
	Social functioning hard to judge (3)	The psychosocial functioning was particularly difficult as we rarely question patients on this aspect of their life. REVIEWER05
	Role functioning hard to judge (3)	I always had to think about the 'role' subscale because I feel that it is such an interpretation of the person answering it. So I always got stuck on that and was never confident about it.REVIEWER15
	Cognitive functioning hard to judge (3)	The ones that were very difficult were financial status and cognitive functioning. REVIEWER38
	Emotional functioning hard to judge (1)	Cognitive –emotional – social more difficult & unpredictable REVIEWER24

Theme (number of interviews extracted from)	Sub-theme	Quotes
	Constipation, diarrhoea, dyspnoea hard to judge (1)	Yes, financial difficulty, constipation, diarrhoea, dyspnoea. ...For the symptoms mentioned above, it is hard to determine whether they were from psychological stress or physical illness. REVIEWER37
	Global health status hard to judge (1)	...global health status is also difficult as relatively insensitive. REVIEWER24
No difference in difficulty (3)		I feel very familiar with the QLQ-C30 and how it works, so no. REVIEWER12
Some subscales easier (4)	Clinical/disease related symptoms easier (4) Fatigue Nausea/vomiting Cough Pain	The easier ones are all the symptoms. Fatigue, nausea and vomiting etc. REVIEWER38 In general I focused on the things that were clinical or disease related like cough or pain. REVIEWER21 It may be easier to make a judgement regarding say a specific treatment end emesis on say the nausea and vomiting scale in some studies. REVIEWER06 the more defined areas like nausea and vomiting which are clearly related to an intervention REVIEWER38
	Global health status (1)	Some like Global health status and fatigue were easier. REVIEWER38

6.1.3.4 Consistency of approach to reviews

Reviewers generally described very similar approaches to the task. They either read the paper first and then the coversheet to determine what comparisons were being judged followed by re-reading the paper or some looked at the coversheet first and then read the paper with the comparisons required already in mind. Most reviewers described going back over their score sheets to check them once they had filled them in, some commenting on how confusing this part was and how carefully it needed completing with the direction and magnitude of expected change. Reviewers reported drawing on their knowledge of patients undergoing similar treatments and consideration of what the QOL scales would show.

Specific sections of the paper mentioned by reviewers as the most useful were research question, patient characteristics, eligibility criteria, description of treatment, methodology, schedule and compliance with QOL forms, primary endpoint results and toxicity (Table 33).

Table 33 Approach to reviews and information used (questions 6/7)

Theme (number of interviews extracted from)	Sub-theme	Quotes
Approach to task described in detail (10)	Read paper before looking at questions (7)	I read the paper a few times first and tried to get an idea of what it was looking at and so on, and then went step by step thought the questions. REVIEWER36
	Read questions before looking at paper (3)	I tended to read the tables for the answers just to get an idea of what groups we were looking at, then I would read the paper and then I'd go back and do the tables. REVIEWER15
	Checked their answers (4)	Then fill in the tables and then go back to check not got it wrong. REVIEWER12
Specified information used from paper (12)	Patient characteristics (8)	Descriptive and demographic details of the sample. REVIEWER12

Theme (number of interviews extracted from)	Sub-theme	Quotes
	Methods/description of treatments (8)	The details about patients' clinical condition and the therapy they underwent were most helpful and useful. REVIEWER37
	Primary/clinical results (3)	..primary results (RR/DFS/OS),.. REVIEWER24
	Research question (2)	I looked at the study design and whether the question they asked was a question that was likely to give a lot of Quality of Life change. REVIEWER38
	Toxicity (1)	toxicity if available REVIEWER24
Other information used to inform decisions (5)	Experience of patients/studies (2)	Then, I go back to my own studies and memories of patient care. REVIEWER37
	Thought about the scale (2)	knowledge of sensitivity of QLQ-C30 to these effects REVIEWER24
	Drug adverse events database (1)	Sometimes I go to the database or information centre for drug adverse effects, etc. of our hospital to collect updated knowledge to help me make decisions. REVIEWER37
	Recognised the study (1)	I tried to recognise it which I did for a couple, and that was interesting because I had prejudged opinions for the *<name of author>* paper as I was familiar with the work. REVIEWER36

6.1.3.5 Time taken to complete the reviews

Averages were reported to be from 10 to 60 minutes. Most reviewers averaged between 15 and 30 minutes per paper.

6.1.3.6 Opinions on experience necessary in order to do the reviews

The answers from questions 9/10 have been combined as there was substantial overlap in the comments (Table 34). Knowledge of cancer and quality of life measurement was required in order to be a reviewer for this project but no minimum level of experience was specified in the invitation letter to reviewers. Seven reviewers agreed that having experience of using the QLQ-C30 in cancer patients was sufficient to do the reviews, while 4 disagreed. A number of the comments received indicate that reviewers feel that both clinical and research experience/interest in the QLQ-C30 is useful and the more experience you have the better able you are to predict. There was a feeling that you needed to know some of the background to the instrument and be familiar with the scores through knowledge of published results or using it in clinical trials as well as using it day to day clinically.

Table 34 Level of experience required to review papers

Theme (number of interviews extracted from)	Quotes
The more experience the better the reviews will be (5)	More experience the more realistic the results. REVIEWER05
Need more in depth knowledge of QLQ-C30 and changes (4)	Think you need to go off and do a bit more reading around the subject to understand what the pitfalls might be. I don't think clinicians using the QLQ-C30 every day would automatically be able to do this. You need to understand some of the background of the instrument. REVIEWER30 I feel it would have been more helpful to have more knowledge of QLQ-C30 about the way patients answer and score various aspects of their treatment. REVIEWER05
Used QLQ-C30 in a disease area (4)	Someone who is used to using the instrument in say a clinical trial will have experience of the sizes of difference. REVIEWER06
Clinical and research experience of QLQ-C30 useful (3)	It might be that people who use the QLQ-C30 in their research have a better handle on it. REVIEWER21 I think that you need more, you need both the clinical experience judging patients not necessarily using the QLQ-C30 as well as experience with QLQ-C30 research REVIEWER36

Theme (number of interviews extracted from)	Quotes
Used QLQ-C30 in a study/studies (3)	yes, at least participating in cross-sectional studies for two different cancers, one for each, better follow-up studies or clinical trials for more cancers. REVIEWER37

6.1.3.7 Opinions on using published QOL data to inform clinical interpretation

10 reviewers thought it was possible to use published data to inform interpretation of QOL scores, 1 was uncertain and 2 did not provide an answer. There was general agreement that published data should be able to be utilised but that it was difficult to do so. Three reviewers suggested improvements/additions to the EBIG methodology and two suggested alternative ways of utilising published data. The themes extracted from answers to this question (question 11) can be found in Table 35.

Table 35 Using published data to inform interpretation

Themes (number of interviews extracted from)	Sub-themes	Quotes
Ideas to improve the EBIG method (3)	Patient involvement (2)	I wonder if it would be interesting to involve patient groups in the review of results involving both patients and relatives. Also, it would be useful to find out if the scores for the Quality of Life surveys agreed with the experience of patient helplines REVIEWER05 I was also thinking of ways to get patients involved. REVIEWER15

Themes (number of interviews extracted from)	Sub-themes	Quotes
	Show reviewers the direction and just get opinions on size (1)	I wonder if you would have got more useful information if some of the blacked out results were given. Not the number or scores, but say if emotional function was improved by treatment, then ask us how significant the change was. Because I think that particularly in longitudinal, when there might have been 4 or 5 time spots, and I know things can go up and down all over the place, the information may not have been useful from this. So sometimes knowing which way it was going it may have got more useful information. REVIEWER15
	Only use high quality studies (1)	First, if you're using published data, you're in the hands of whoever did the study to how well it was done and how honestly it was done. I would only accept published data if the group was reputable and I was really confident that they had put in the manuscript exactly how they had gone about obtaining the Quality of Life data, for example many people failed to complete the forms. I think the longer the study goes on Quality of Life changes become less valuable. REVIEWER38

Themes (number of interviews extracted from)	**Sub-themes**	**Quotes**
Other ideas for utilising published data (2)	Reference values (2)	Yes, as a reference value. E.g. for a certain treatment of a certain cancer, how much a certain scale will change or differ. Physicians or nurses can take these references for their own care plans and even care quality evaluation. REVIEWER37 It must be possible to use data to inform clinical interpretation, but it is a matter of presenting that data in a way a clinician will understand. I guess the whole point of the project is to get a yard stick from which to work from. I think the only way to do that is to relate it to some other yard stick that a clinician would understand such as the Karnofsky performance score, e.g. the difference in QLQ between KPS (*Karnofsky Performance Status Scale*) 100 and KPS 80 is X. Then clinicians could identify the size of QOL difference related to something. REVIEWER21
	Use the discussion from the paper to interpret changes (1)	Tricky but, yes. You could look at just the discussions and pick out the narrative and contextual description of the impact on QOL, if the clinicians who have written papers have put into words exactly what the results really mean. REVIEWER12
General comments (5)	Should be able to utilise it but unsure how (3)	If you can't use published data though, what's the point of it being published at all? You must be able to utilise it somehow, but I don't think we yet know the best way to utilise it. REVIEWER30
	Uncertain if it can be utilised (1)	Hard to tell, based on guessing, very uncertain way of doing it. REVIEWER04

Themes (number of interviews extracted from)	Sub-themes	Quotes
	Useful for group interpretation not individual (2)	I suppose it depends on what you want to be able to do though. If you are wanting to feed it back to patients, or to help an individual patient, or if you have results of a clinical trial and you want to know whether it is a meaningful result as the result of the trial. REVIEWER30 The trouble is that published data is derived from group data, and the individual ones are individual. But yes, I think there has to be a starting point, but you may have to refine it. REVIEWER22
	EBIG a good idea that works (1)	I think that the idea is great and that it does work REVIEWER15

6.1.3.8 Additional comments

A number of reviewers commented on how useful and interesting they had found the project (see Table 36). There was interest in seeing how other reviewers had approached the task and in seeing the results. Reviewers also remarked on how educational the task had been as it had made them read around the subject more and think more about patient perspectives.

One reviewer in a disease area with very few papers pointed out that all the published papers were using the same treatment therefore results may not be very generalisable in that disease area. They also pointed out that these papers were very well known to them. They also felt that the QLQ-C30 in that disease area may not be detailed enough to measure the QOL of these patients.

Table 36 Additional comments

Themes (number of interviews extracted from)	Quotes
Interesting to do (4)	This was interesting to do and it made me think about the patient perspective on treatment etc. REVIEWER05
Possible issues with experts who are not medical doctors (2)	I am not familiar with the intricacies of each type of treatment as for example a medical doctor would be. So I found that quite difficult REVIEWER36
Possible influence of language/cultural differences (1)	Cultural difference or misunderstanding due to language difference may also affect my guess. REVIEWER37
Issues specific to brain cancer (1)	For brain as so few papers and all are based on a single drug, in many ways it is limited. Very well tolerated drug. For the brain part this may not be a very representative/generalisable method. REVIEWER04
Guessing (1)	Sometimes I felt the review was like a 'guess'. If I can have more information, my 'guess' will be more accurate. REVIEWER37
Direction of difference confusing on judging scale (1)	...found the way the cover sheets were written counterintuitive, where I could have just written this shows x, y, or z, I found filling in whether it was minus 3 or plus 3 very difficult, and I was worried I'd got it wrong and gone the wrong way. REVIEWER12

6.1.4 Conclusions

Feedback was not received from all the reviewers so it is important to note that this may be a biased sample, consisting only of those reviewers who were willing to spend time taking part in a telephone interview or who requested the questions via mail instead.

One of our concerns about the project was that reviewers may already have clear ideas of a clinically relevant difference based on previous work, e.g. 10 points, and that

the EBIG reviews would be based on these rather than familiarity with the scores. Although reviewers may have had an idea of clinical relevance it is encouraging to note that there is no indication that their reviews involved guessing what the scores might be and then using existing estimates of clinical relevance to make their judgements.

Reviewers generally approached the reviews in a similar way indicating that the guidelines provided for reviewers were adequate in the description of the task.

The interviews with the experts were also useful in identifying how, if this methodology was to be repeated for other QOL measures, experts should be defined. As this was the first time a panel of this nature had been convened the definition of an expert was left rather vague. Since we required experts for a wide range of cancer types and contrasts we only specified that an expert should have experience of cancer and its treatment as well as the EORTC QLQ-C30. These interviews have highlighted that experience of the EORTC QLQ-C30 in just a research setting or just a clinical setting may be too narrow for a reviewer to judge papers from both settings. It may be that experts who have research experience with the QOL instrument as well as clinical experience can conduct the reviews more easily and accurately. Experts' familiarity with treatments is important in their perceived ability to judge the papers. Reviewers with narrow experience appeared to find the task more difficult. If repeated the recommendation would be to match the papers more closely to the experience of the experts, in terms of the setting (research/clinical) and according to familiarity of treatments/interventions in each setting. However, this may not be feasible as it would mean an even larger panel of reviewers was required and each reviewer would undergo training in order to then review only a very small number of papers. A better solution may be for papers containing relatively unfamiliar treatments/interventions to be excluded. Stricter criteria may be required for experts conducting these reviews to ensure that they do not judge papers containing unfamiliar treatments or clinical settings. Although it was an option for reviewers to return any papers without judgement if they felt unable to comment it may be that some experts chose to guess instead which may create more variation in the results and could be a cause for discrepancies between reviewers of the same papers.

Instead of obtaining expert opinion on the size and direction of the difference in QOL we could have given the reviewers an indication of the direction of the change in order to concentrate the reviews on size only. This was suggested in one of the interviews. One advantage of this over the review process adopted would be that discrepancies between reviewers assuming different directions in QOL change would

have been eliminated and with only the magnitude being considered the review process would be simplified for the reviewers. We undertook a data cleaning exercise for papers where reviewers' opinions were in opposing directions but a number of discrepancies remained. One disadvantage of indicating direction is that the 'no difference' category may be under-used as an indication of direction points to some change in QOL scores. The major advantage of masking the direction as well as magnitude (as carried out in this project) is that comparisons where experts cannot agree on the direction of their judgements are highlighted as potentially comparisons that are not familiar enough to the experts or the anchor is not closely related enough to QOL to be used to inform clinical interpretation of QOL scores.

Two reviewers quite rightly suggested patient involvement would be useful in production of the interpretation guidelines and the results from our patient pilot study can be found in Chapter 8.

Two reviewers suggested an alternative way of using published data was as reference values. The main issue in trying to use the published data in this way would be the availability of data from a number of papers on the same anchor. It is unlikely that enough data is published at the level of specific treatments within cancers but it is likely that this is possible for other widely used anchors, e.g. ECOG performance status. If there are a number of papers reporting QOL scores for particular anchors and it appears reasonable from the analysis that results can be pooled over cancer types then reference values can be published in these interpretation guidelines alongside the more general guidance.

One reviewer suggested a different way of utilising published data for interpretation would be to just use the discussion and interpretation from the paper on the meaning of the QOL changes. Following a review of the RCTs identified in the literature search for this project only 38% of studies include a clinical interpretation of the QOL scores (3) so this method would only be able utilise some of the published data. Our review found that there was generally an over-reliance on statistical results rather than clinical interpretation of differences in QOL.

One reviewer highlighted that the quality of the reviews is reliant on the quality of the underlying paper and how the study was conducted. This is a key issue with using published data to produce the interpretation guidelines and the second part of this Chapter investigates what constitutes a 'good quality' paper for this meta-analysis and how to weight the papers according to quality in the analysis.

Some of the reviewers' comments highlight areas that may affect the quality of the review:-

a) study design and complexity
b) heterogeneity in patient population
c) attrition rate
d) anchors unrelated to QOL
e) certain subscales which were harder to judge

All these factors are considered further in Part 2 of this chapter investigating the quality of the expert reviews and the possible effect of quality on the overall study results.

6.1.5 Discussion

One aim of the interviews was to find out how experts approached the task and to see how consistent the panel of experts were. We found that despite having a large panel of reviewers working independently, the interviews revealed that the reviewers generally took a consistent approach to the reviews (Table 33). Reviewers highlighted patient characteristics and the methods or treatment descriptions as the main sources of information from the masked papers to inform their reviews.

Although experts were generally familiar with different literature on the size of meaningful QOL differences it was encouraging to also note that there was no indication that this dominated the approach to the reviews.

The interviews also sought to highlight any issues with particular papers, subscales or comparisons which more generally may affect the quality of the expert reviews and subsequently the results. A number of points were raised which were useful in guiding sensitivity analyses and investigations of quality in the meta-analysis. Contrasts with unfamiliar treatments/interventions, heterogeneous patient groups, high attrition rate and complex trials (e.g. multi-stage trials) were generally harder for experts to judge. Anchors highlighted as difficult were time, side effects, those which the reviewer felt had little relationship to QOL and contrasts comparing groups fairly close together (e.g. ECOG 0 versus ECOG 1 compared with ECOG 0 versus ECOG 4). These may be factors that could affect the quality of the reviews in terms of the concordance between reviewers' opinions or agreement between the reviewers' opinions and the study results. These contrasts may also be the ones with the most uncertainty in reviewers' opinions (i.e. percentages spread across the possible categories). Using a weighted average of reviewers' opinions in these cases may lead

to more variation in the pooled results. This is investigated quantitatively in Part 2 of this chapter.

There was no indication that either longitudinal or cross-sectional contrasts were easier for experts to judge.

Financial difficulties, social, role and cognitive functioning were harder for experts to judge than subscales which are more symptom-related.

The definition of an expert for the project may have been too non-specific. A wide range of experience (e.g. clinical and research experience) with the EORTC QLQ-C30 may be necessary in order to be familiar enough with how the scores behave in different situations. It may be that experts from different backgrounds or with experience only in the clinic or in research had different opinions on the likely size of difference, leading to discordance between experts. However, this could also be a strength of our choice of panel in that a wider range of views was sought and an average of these was then used to class the contrasts.

These qualitative interviews were valuable in identifying factors that may influence how an expert judges the size of a difference in QOL. A number of these factors would not have been hypothesised without undertaking the interviews. These factors are used in the next chapter to explore concordance between reviewers on the same contrast and uncertainty in expert judgements quantitatively. This was with the aim of understanding the quality of the expert opinions and how the results from the project may be affected by including papers or contrasts the reviewers found hard to judge.

At the start of the project we aimed to train the experts and provided a manual(68) to ensure they approached the reviews in the same way and these interviews are reassuring in that they show this appears to have been successful.

6.2 Part 2: Meta-analysis quality assessment

6.2.1 Correlation between reviewers scores and actual QOL changes

6.2.1.1 Cross-sectional versus longitudinal contrasts

Different measures for measuring correlation (parametric and non-parametric) gave almost identical results therefore Pearson's correlation coefficients are used here for simplicity. Correlation between the reviewers' and actual QOL scores was 0.31 for the cross-sectional contrasts (Figure 13) and 0.21 for the longitudinal contrasts (Figure 14). Note the graphs show individual reviewers' weighted averages (individual opinion) at this point rather than the average for a contrast (overall opinion) so this contains multiple observations on the same contrast. Note also that the randomised comparisons at baseline which were not sent for review were excluded here in order to get a picture of how well the experts were correlating with the actual scores rather than looking at the dataset as a whole. Although there is an overall positive trend for both types of contrasts (i.e. showing a positive relationship between the reviewers' weighted averages and the scores from the original article) the scatterplots show there is a lot of variation. There are a number of sizeable mean differences falling in the trivial to small range and, at the other ends of the scale, a number of very small mean differences in the medium to large range.

Figure 13 Correlation between reviewers' scores and actual QOL scores – cross-sectional contrasts

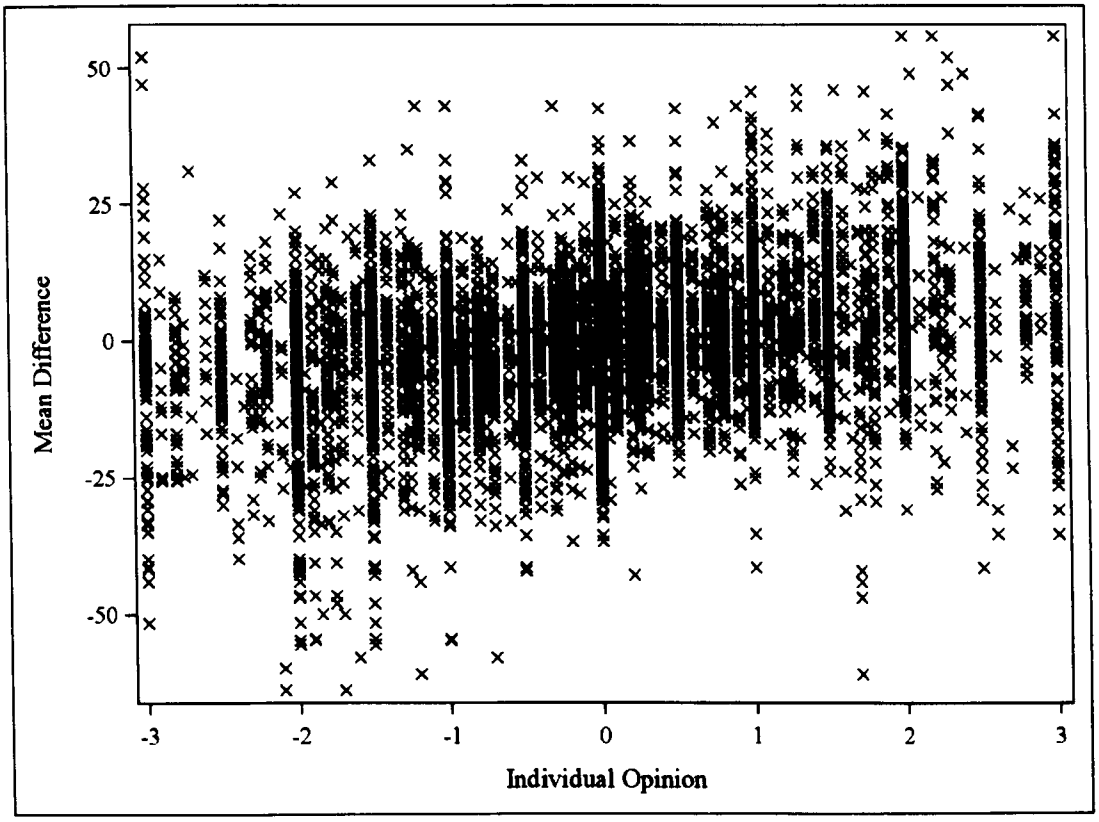
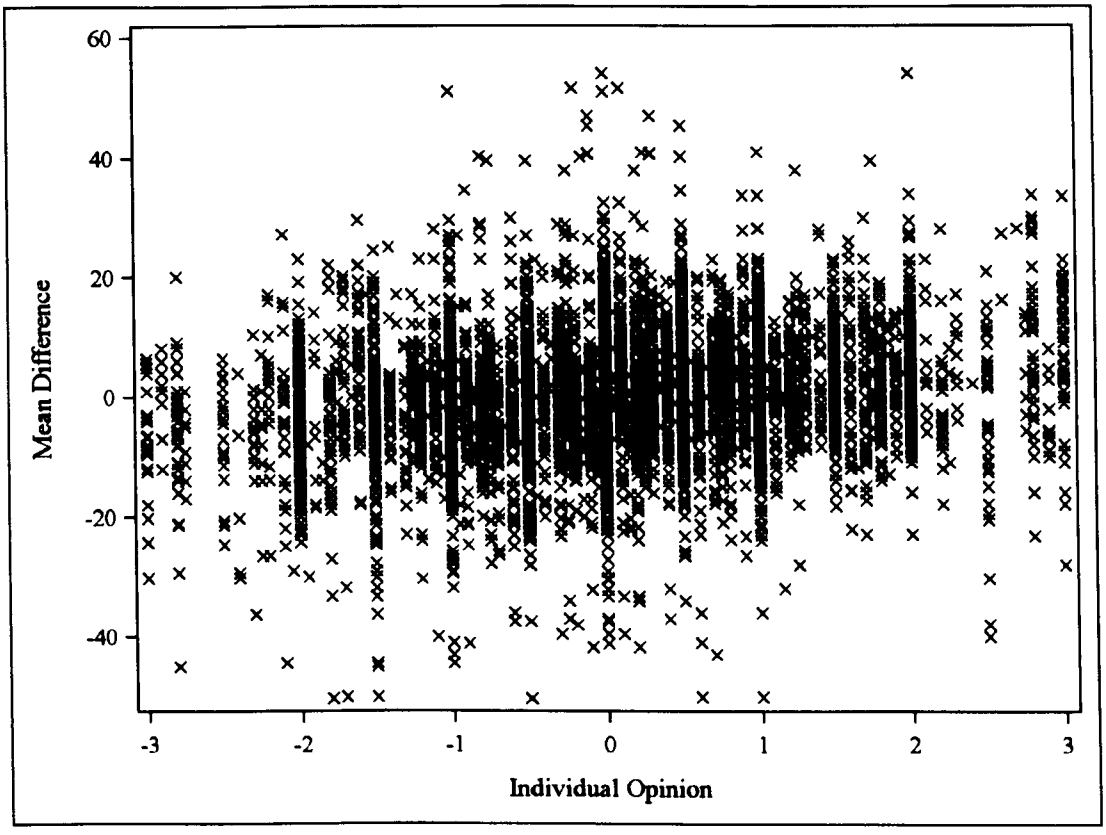


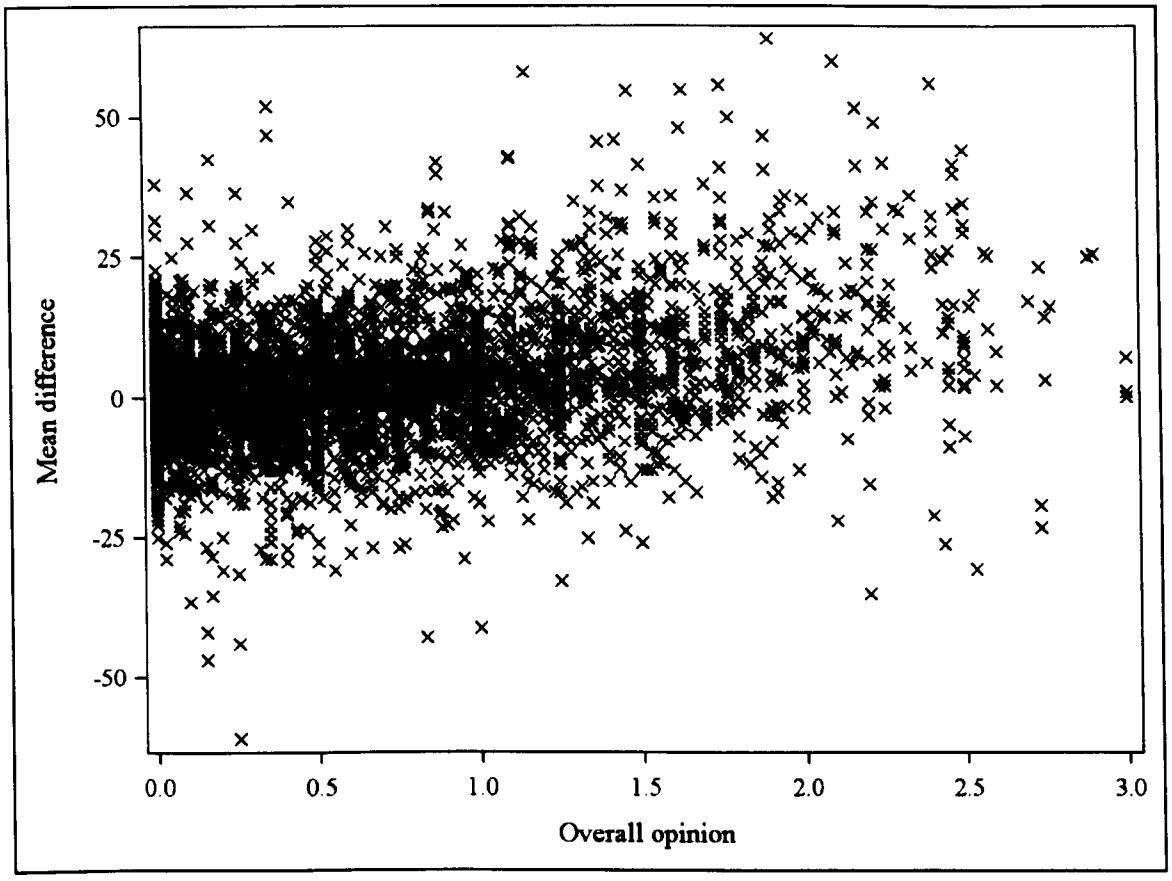
Figure 14 Correlation between reviewers' scores and actual QOL scores – longitudinal contrasts



I also looked at the relationship between the overall weighted averages (i.e. on a contrast level rather than on an individual review level) and the actual mean scores as this is more relevant as a measure of correlation between the grouping variable we used in the analysis and the paper scores. Since we are now looking at the relationship overall between the size classes and actual QOL scores the contrasts not sent for review are now included. The negative and positive reviews were also combined here as they were for the analysis (i.e. contrasts with a negative review have the expert scores and mean differences multiplied by minus 1).

Figure 15 shows the correlation between the mean difference and overall opinion for the cross-sectional contrasts. The Pearson's correlation coefficient is 0.32 ($p<0.0001$), which is similar to that when considering the individual opinion (0.31).

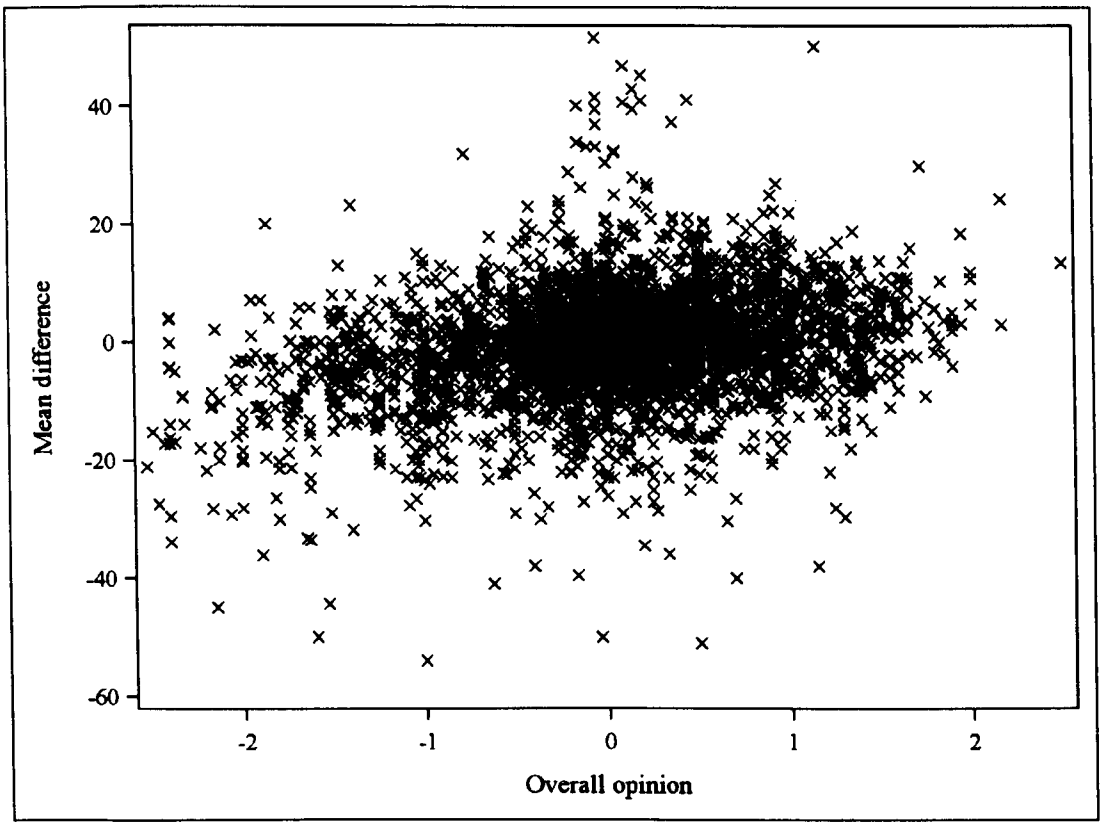
Figure 15 Relationship between overall opinion and actual QOL scores (cross-sectional contrasts)



For the longitudinal contrasts (Figure 16) the correlation is 0.28 ($p<0.0001$), although the majority of contrasts are around the trivial and small categories (-1 to 1)

the positive relationship between the average expert score and actual QOL difference can still be seen.

Figure 16 Relationship between overall opinion and actual QOL scores (longitudinal contrasts)



6.2.1.2 Individual subscales

The scatter plots in Appendix II show the overall opinion against the mean difference from the paper for each subscale. The correlations are summarised in Table 37. Longitudinal and cross-sectional contrasts are combined in these analyses. Across subscales the correlations ranged from -0.05 to 0.47. PF and FA had correlations of >0.4. The subscales showing very little correlation (<0.2) between the overall opinion and the paper mean difference were SL, DY, EF and FI.

Table 37 Correlation of reviewers' scores with actual QOL differences (by subscale)

Subscale	Pearson's correlation	Number of contrasts
PF	0.47	636
FA	0.42	578
NV	0.39	478
RF	0.38	606
QL	0.36	711

Subscale	Pearson's correlation	Number of contrasts
AP	0.35	462
CF	0.33	541
PA	0.32	533
SF	0.30	648
CO	0.25	416
DI	0.24	406
DY	0.16	404
SL	0.16	451
EF	0.15	627
FI	-0.05	336

6.2.1.3 Individual reviewers

Plots of the relationship between reviewers' weighted averages and actual QOL differences can be found in Appendix III. Plots were produced for each reviewer in order to see if there were particular reviewers that stood out as outliers from the rest. For individual reviewers correlations range between -0.24 and 0.64 (Table 38). Correlations are highlighted using italics in the table below where the p-value indicates the correlation is not significantly different from zero or where a significant negative correlation exists between the reviewer and actual scores (as this indicates the reviewer tended to judge in the opposite direction to the paper). $P<0.01$ for all other reviewers indicating a significant correlation between their judgements and the QOL differences in the paper. Reviewers 4, 6, 8, 10, 16, 20 and 36 are highlighted as those with poor correlation of their judgments with QOL differences in the paper. These reviewers all reviewed a small number of papers (five or less papers each), with the exception of reviewer 36 who reviewed 12 papers in total but still had a correlation of close to zero.

Table 38 Correlation of reviewers' scores with actual QOL differences

Reviewer ID	Number of papers	Number of contrasts	Pearson's correlation	P-value
2	2	21	0.64	0.002
25	10	310	0.53	<0.0001
17	8	347	0.51	<0.0001
26	15	410	0.50	<0.0001

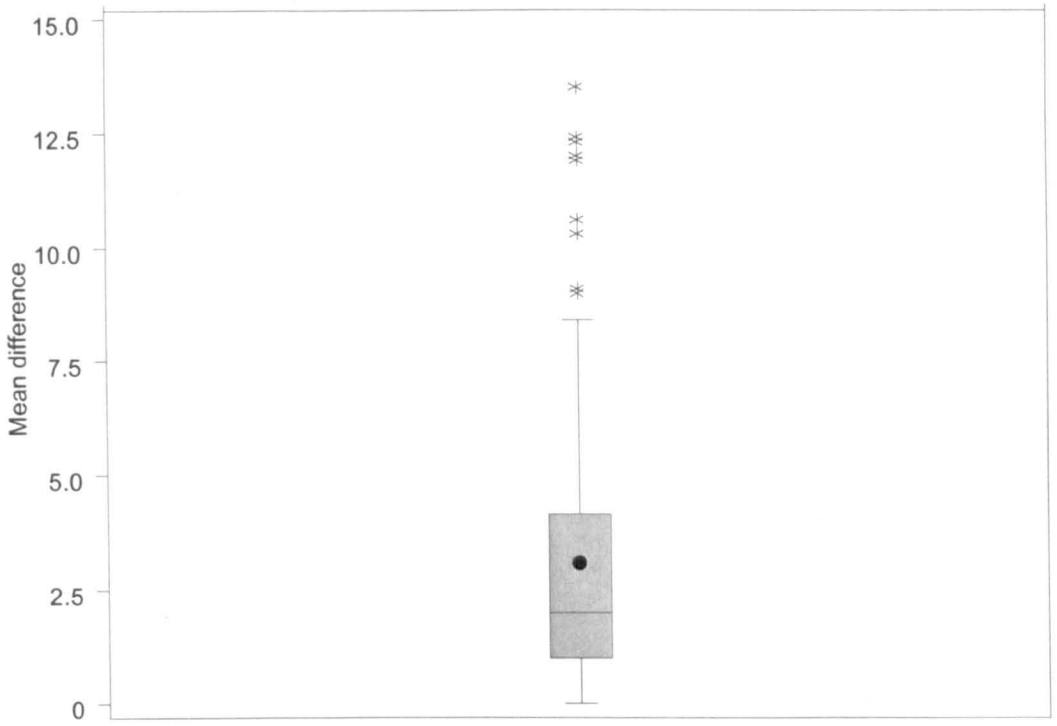
Reviewer ID	Number of papers	Number of contrasts	Pearson's correlation	P-value
5	19	487	0.45	<0.0001
24	45	1502	0.44	<0.0001
35	4	66	0.43	0.0003
22	2	96	0.42	<0.0001
15	29	1229	0.41	<0.0001
14	15	434	0.41	<0.0001
3	8	244	0.36	<0.0001
13	2	57	0.35	0.008
31	19	392	0.34	<0.0001
1	26	722	0.33	<0.0001
19	46	1471	0.33	<0.0001
29	49	1585	0.32	<0.0001
21	28	1236	0.31	<0.0001
9	11	382	0.30	<0.0001
7	25	550	0.30	<0.0001
30	19	595	0.27	<0.0001
12	31	780	0.25	<0.0001
37	43	1561	0.23	<0.0001
11	70	2563	0.21	<0.0001
18	4	160	0.20	0.02
23	6	422	0.17	0.0006
16	*1*	*30*	*0.16*	*0.39*
28	10	491	0.14	0.002
27	16	629	0.11	0.007
8	*4*	*189*	*0.07*	*0.35*
36	*12*	*307*	*-0.01*	*0.92*
4	*4*	*190*	*-0.03*	*0.65*
10	*1*	*75*	*-0.09*	*0.43*
20	*1*	*105*	*-0.21*	*0.04*
6	*5*	*265*	*-0.24*	*0.0001*

6.2.1.4 Subset of trivial contrasts not sent to reviewers

The comparisons between randomised treatment groups but at the baseline time we thought should be automatically assigned to the trivial size class. These were not

sent for expert review therefore I have also summarised here the actual sizes of QOL differences that arose from these contrasts. They had a mean difference of 3.1 points (median 2.0) and a range from 0 to 14. 97% of the values were less than 10 points (Figure 17). The outlying points with mean differences above 10 points were all checked for data entry errors and were found to be genuine differences at baseline.

Figure 17 Distribution of actual mean differences for contrasts not sent to reviewers



6.2.2 Concordance between reviewers on the same contrasts

6.2.2.1 Distance between reviewers

Seventy-six percent of the cross-sectional contrasts had a maximum distance between reviewers of up to one size class (Figure 18). This was lower for the longitudinal contrasts (63%), Figure 19. For the cross-sectional contrasts there were 35% in exact agreement. There were less in complete agreement for the longitudinal contrasts (21%). There were very few contrasts a distance of three or more size classes apart (8%).

Figure 18 Distance between reviewers for cross-sectional contrasts

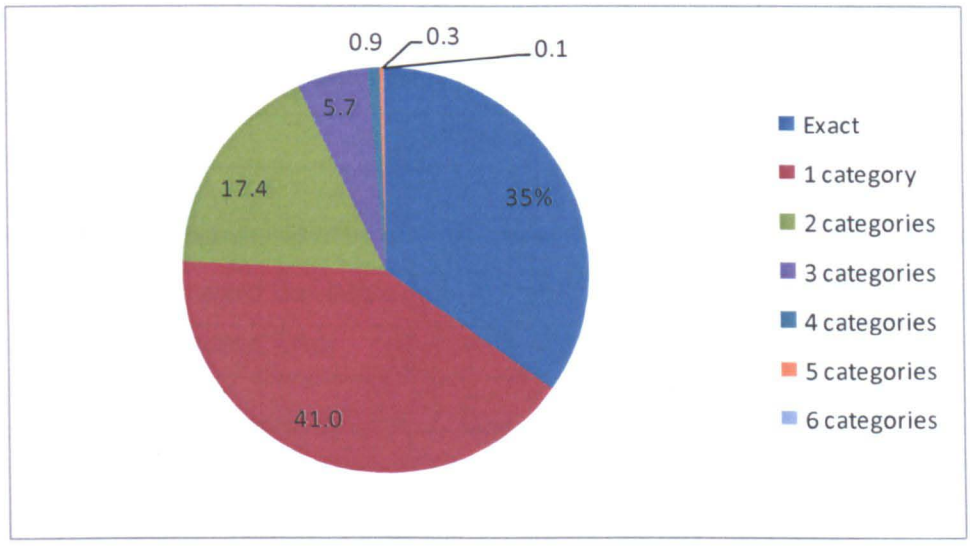
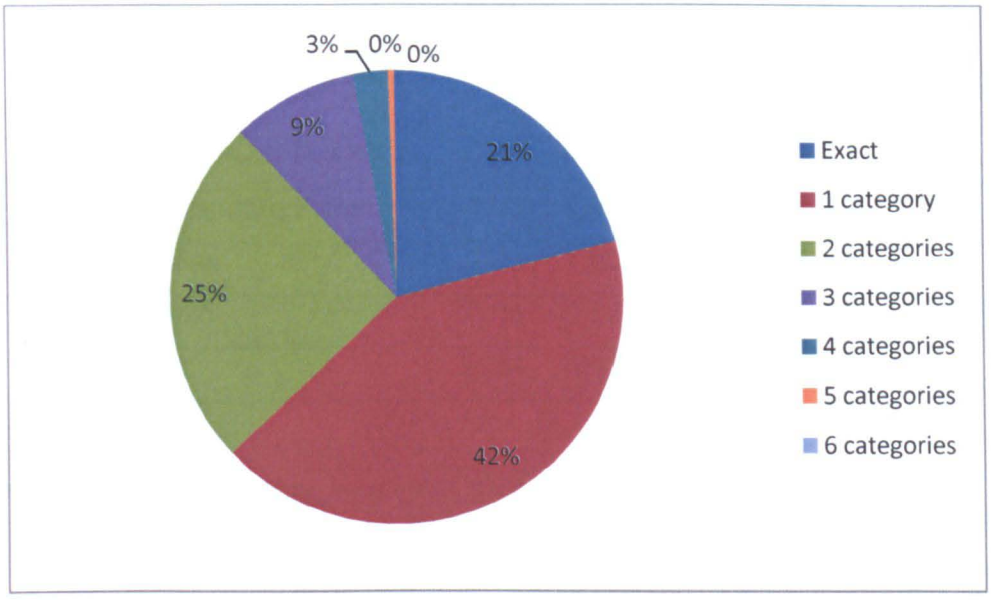


Figure 19 Distance between reviewers for longitudinal contrasts



6.2.2.2 Consensus measure

The average consensus score was 0.90 for the cross-sectional contrasts and 0.87 for the longitudinal contrasts (Table 39). The examples I showed in the Methods chapter (Table 13) show that a value of around 0.9 is achieved when the reviewers weighted averages are in adjacent categories. Summarising across expert size classes (Table 40) shows that the highest agreement is for cross-sectional contrasts is for the trivial differences (even removing those not actually sent for review). For the longitudinal contrasts the consensus is similar regardless of the expert size class

(Table 41). The consensus scores were also summarised by subscale (data not shown) but there were no noticeable differences across subscales.

Table 39 Consensus score by contrast type

	Cross-sectional	Longitudinal
Mean	0.90	0.87
Median	0.90	0.89
Standard Deviation	0.09	0.10
Standard Error	0.00	0.00
Max	1.00	1.00
Min	0.35	0.43
N	4215	3397

Table 40 Consensus score by expert size class (cross-sectional contrasts)

	Trivial	Small	Medium	Large
Mean	0.94	0.86	0.89	0.93
Median	1.00	0.89	0.89	0.90
Standard Deviation	0.08	0.10	0.08	0.05
Standard Error	0.00	0.00	0.00	0.01
Max	1.00	1.00	1.00	1.00
Min	0.36	0.35	0.65	0.89
N	2034	1696	450	35

Table 41 Consensus score by expert size class (longitudinal contrasts)

	Deterioration				Improvement		
	Large	Medium	Small	Trivial	Small	Medium	Large
Mean	0.90	0.86	0.87	0.89	0.86	0.85	0.89
Median	0.90	0.89	0.89	0.90	0.89	0.89	0.89
Standard Deviation	0.00	0.08	0.09	0.10	0.09	0.10	.
Standard Error	0.00	0.01	0.00	0.00	0.00	0.01	.
Max	0.90	1.00	1.00	1.00	1.00	1.00	0.89
Min	0.90	0.57	0.43	0.51	0.51	0.65	0.89
N	2	148	543	1774	850	79	1

6.2.2.3 Between-reviewer SD

The average between-reviewer standard deviation was 0.5 (Table 42) for the cross-sectional contrasts and 0.7 for the longitudinal contrasts (Table 43). The distribution of the between-reviewer SD was positively skewed for both types of contrast (Figure 20 and Figure 21), i.e. the bulk of the values lie to the left of the mean, towards zero. Note the proportion of contrasts with zero SD will not be the same as the proportion with zero distance between reviewers shown in 6.2.2.1, since the former rounds the weighted averages to the nearest size class.

Table 42 Between-reviewer standard deviation – cross-sectional contrasts

Subscale	Mean	Median	SD	SE	Max	Min	N
AP	0.48	0.36	0.49	0.03	3.04	0.00	249
CF	0.48	0.35	0.46	0.03	2.12	0.00	294
CO	0.47	0.35	0.49	0.03	2.09	0.00	228
DI	0.39	0.14	0.49	0.03	1.98	0.00	217
DY	0.45	0.32	0.49	0.03	2.33	0.00	213
EF	0.56	0.52	0.45	0.02	2.83	0.00	344
FA	0.59	0.51	0.49	0.03	3.75	0.00	292
FI	0.48	0.35	0.52	0.04	2.98	0.00	196
NV	0.52	0.47	0.48	0.03	1.83	0.00	248
PA	0.55	0.48	0.54	0.03	3.75	0.00	288
PF	0.60	0.53	0.47	0.03	2.84	0.00	336
QL	0.62	0.57	0.45	0.02	2.62	0.00	390
RF	0.59	0.53	0.44	0.02	2.47	0.00	328
SF	0.54	0.50	0.44	0.02	2.52	0.00	354
SL	0.51	0.42	0.49	0.03	2.43	0.00	238
All	0.53	0.48	0.48	0.01	3.75	0.00	4215

Figure 20 Distribution of between-reviewer SD – cross-sectional contrasts

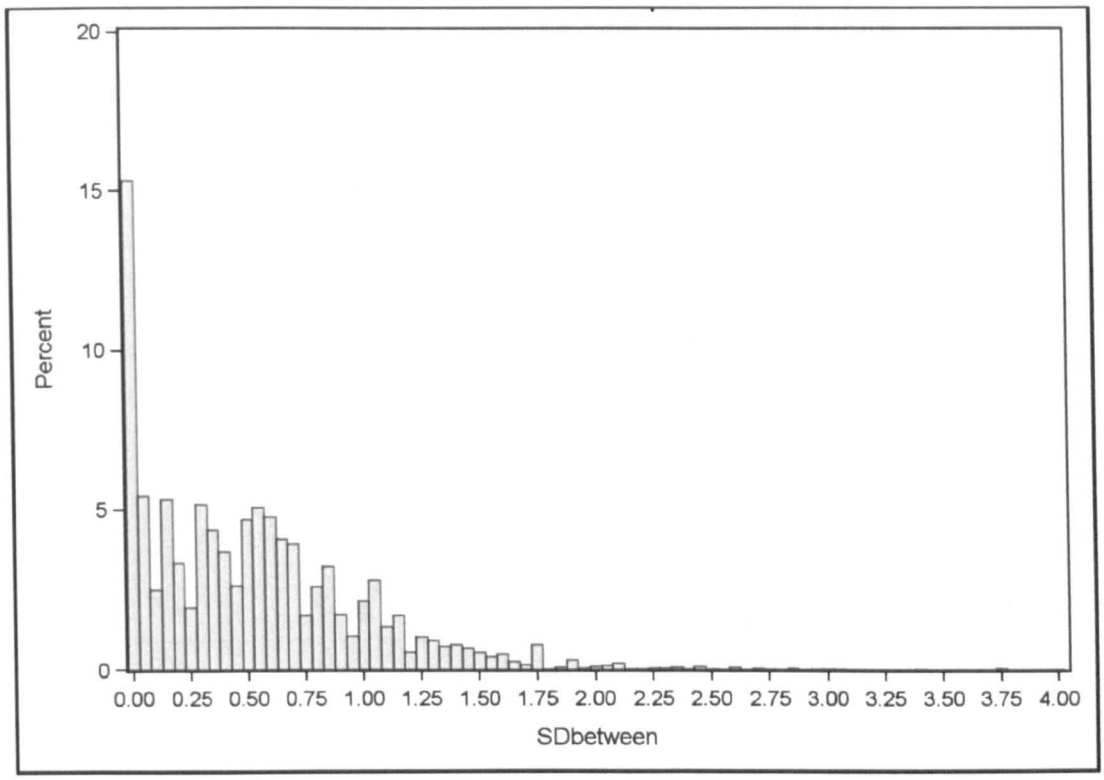
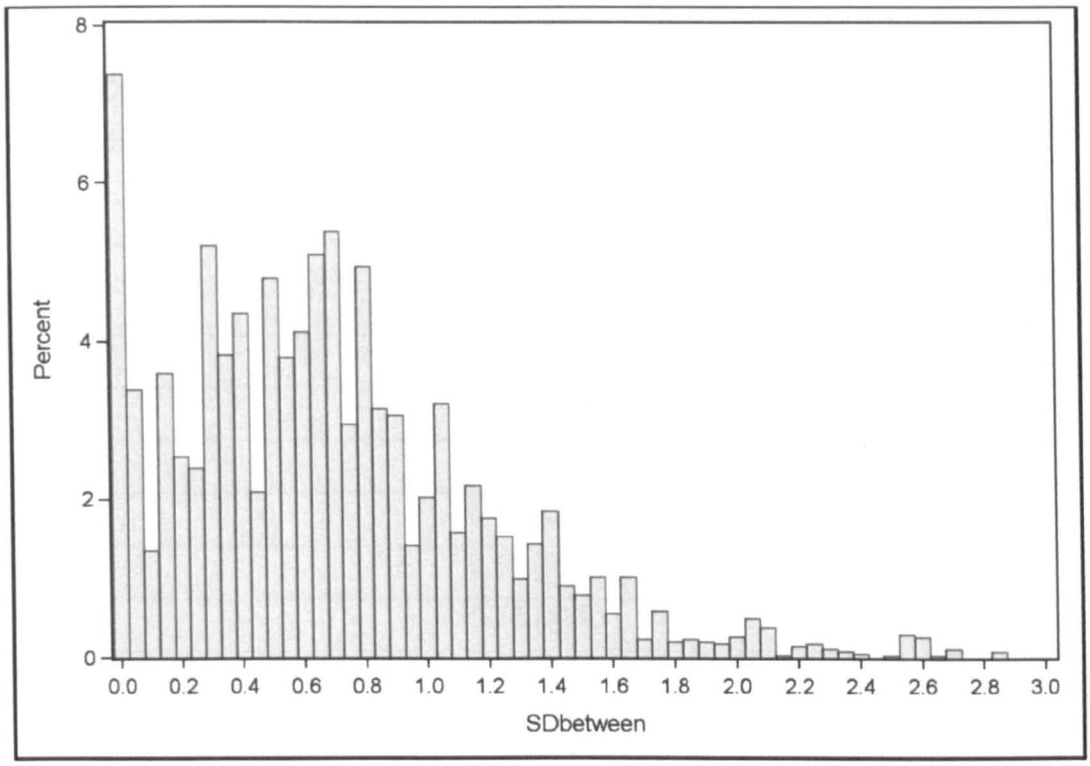


Table 43 Between-reviewer standard deviation – longitudinal contrasts

Subscale	Mean	Median	SD	SE	Max	Min	N
AP	0.71	0.64	0.54	0.04	2.37	0.00	199
CF	0.57	0.49	0.51	0.03	2.62	0.00	231
CO	0.52	0.42	0.48	0.04	1.98	0.00	175
DI	0.47	0.40	0.46	0.03	1.91	0.00	176
DY	0.58	0.50	0.45	0.03	1.67	0.00	177
EF	0.74	0.64	0.52	0.03	2.83	0.00	267
FA	0.78	0.76	0.49	0.03	2.62	0.00	268
FI	0.58	0.50	0.56	0.05	2.67	0.00	130
NV	0.66	0.63	0.48	0.03	2.55	0.00	215
PA	0.72	0.67	0.47	0.03	2.37	0.00	230
PF	0.80	0.71	0.50	0.03	2.69	0.00	284
QL	0.86	0.78	0.55	0.03	2.83	0.00	304
RF	0.79	0.65	0.51	0.03	2.62	0.00	263
SF	0.75	0.67	0.51	0.03	2.62	0.00	279
SL	0.60	0.57	0.42	0.03	1.74	0.00	199
All	0.70	0.64	0.51	0.01	2.83	0.00	3397

Figure 21 Distribution of between-reviewer SD – longitudinal contrasts



6.2.2.4 A note on ICCs

The ICCs were planned in order to use a familiar measure to assess the concordance. However, there is an issue with ICCs when the scale of measurement is small. This is discussed further in light of the ICC results below.

ICCs are displayed for the cross-sectional and longitudinal contrasts separately (Table 44) and for each subscale. ICCs for the cross-sectional contrasts ranged from 0.18 (FI subscale) to 0.6 (PF subscale). CO, DI and FI have ICCs less than 0.4 ('poor' agreement). The remaining subscales would be considered 'fair' agreement, with the exception of PF which falls in the category of 'good' agreement.

ICCs for the longitudinal contrasts were generally lower, ranging from 0.06 (FI subscale) to 0.37 (NV subscale). All subscales would be considered as 'poor' agreement.

The implication of the ICCs seems to be at odds with that from the consensus scores and the high proportion of contrasts we found to have either perfect agreement or maximum differences of only one size class. This is an issue with using ICC. The ICC is dependent on the range of the measurement(88), the greater the variation in the measurement, the greater the ICC. Here we were only measuring on a scale of -3 to 3 (and the extremes of the scale were rarely used by reviewers so in reality the scale is

even smaller). Therefore our ICCs look very small when in fact we know that the agreement is much better than the ICC might imply.

Table 44 ICCs for the full dataset – by subscale and comparison type

Subscale	Cross-sectional contrasts ICC (no of contrasts)	Longitudinal contrasts ICC (no of contrasts)
AP	0.48 (249)	0.20 (199)
CF	0.50 (294)	0.16 (231)
CO	0.35 (228)	0.31 (175)
DI	0.23 (217)	0.34 (176)
DY	0.43 (213)	0.20 (177)
EF	0.56 (344)	0.23 (267)
FA	0.55 (292)	0.35 (268)
FI	0.18 (196)	0.06 (130)
NV	0.41 (248)	0.37 (215)
PA	0.50 (288)	0.28 (230)
PF	0.60 (336)	0.30 (284)
QL	0.57 (390)	0.25 (304)
RF	0.57 (328)	0.30 (263)
SF	0.58 (354)	0.30 (279)
SL	0.41 (238)	0.28 (199)

6.2.3 Factors affecting concordance

The $SD_{between}$ as defined in section 4.4.3.4.2 was used as the outcome variable to investigate which factors may be associated with improved concordance. Note that a high $SD_{between}$ indicates more discordance and a low $SD_{between}$ indicates better concordance.

Boxplots were used initially to visualise how $SD_{between}$ varied across the categories of each factor. The boxplots use a '+' symbol to indicate the mean and a line across the box for the median value. The box represents the inter-quartile range and the whiskers represent the minimum and maximum values. See Table 14 for the key to the abbreviated labels in the plots.

Tables were used to show the results from mixed models used to investigate the significance of each factor. In the table, an estimate with a '+' sign this indicates a higher $SD_{between}$ (more discordance) compared with the reference category for that factor, whereas a '-' sign indicates a lower $SD_{between}$ or better concordance.

6.2.3.1 Factors affecting concordance (cross-sectional contrasts)

Each factor is discussed below. Boxplots showing the distribution of $SD_{between}$ for each level of the factors can be found in Figure 22 to Figure 27. The results from the mixed models can be found in Table 45.

For the study design factor phase III RCTs were used as the reference category and all other designs compared to the concordance for the contrasts from phase III studies. Contrasts from the cohort studies were the only contrasts varying significantly from the phase III contrasts in terms of concordance. The cohort contrasts had significantly more discordance ($p<0.0001$).

For the disease factor, breast cancer was chosen as the reference category and all other cancer types compared to contrasts from breast cancer studies. A number of the other cancer types differed significantly from breast cancer with respect to concordance. Lung, mixed cancer types and testicular cancer were similar to breast cancer. Brain, colorectal, GI, H&N and prostate had significantly higher discordance ($p<0.0001$ to $p=0.0027$). Haematology and urology/kidney had significantly better concordance ($p=0.0008$ and $p<0.0001$ respectively).

The types of anchor all differed significantly from the reference category (treatment-related anchors), all with increased discordance ($p<0.0001$ to $p=0.01$). Time related anchors (e.g. contrasts such as; active treatment group versus follow-up group, <1yr post BMT versus >1yr post BMT) had the worst concordance.

Contrasts of patients with a mix of early and late disease stage had significantly higher discordance compared to those contrasts from early disease patients ($p<0.0001$). Contrasts from late disease stage patients also had slightly higher discordance than the early stage contrasts ($p=0.01$).

There were a number of contrasts where the timing of the comparison could not be identified from the full article. These contrasts with unknown timing had significantly higher discordance when compared with the contrasts at baseline ($p<0.0001$). There was no difference in concordance when comparing the baseline contrasts with contrasts post-baseline ($p=0.03$).

Where an anchor was considered to be well-known the concordance was highest. For contrasts we considered to be known to some experts but unfamiliar to others (variable) and for contrasts where we were unsure of their relevance there was significantly higher discordance ($p<0.001$).

Figure 22 Between-reviewer SD by study design

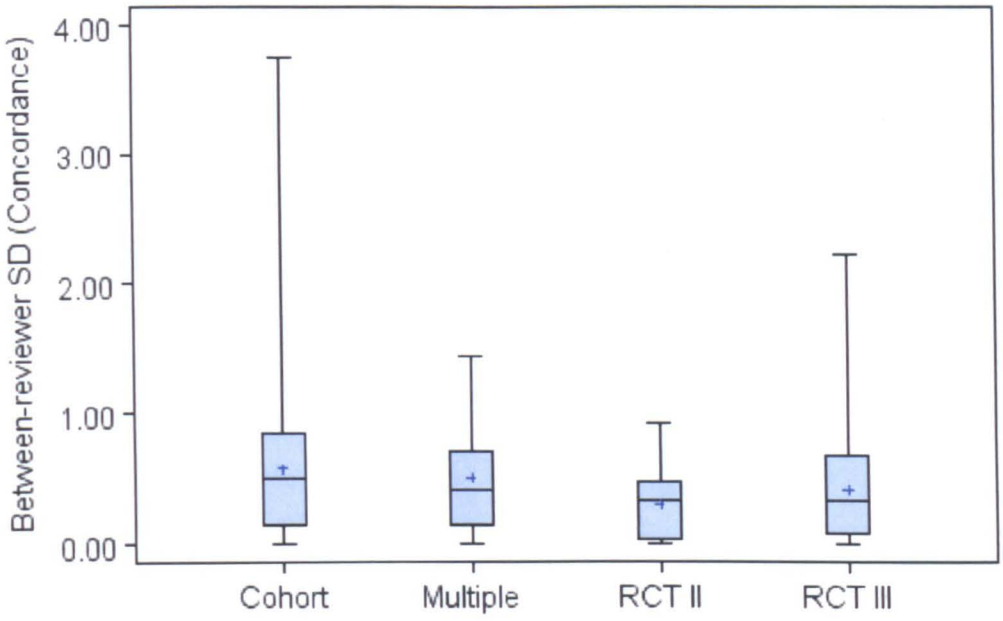


Figure 23 Between-reviewer SD by cancer type

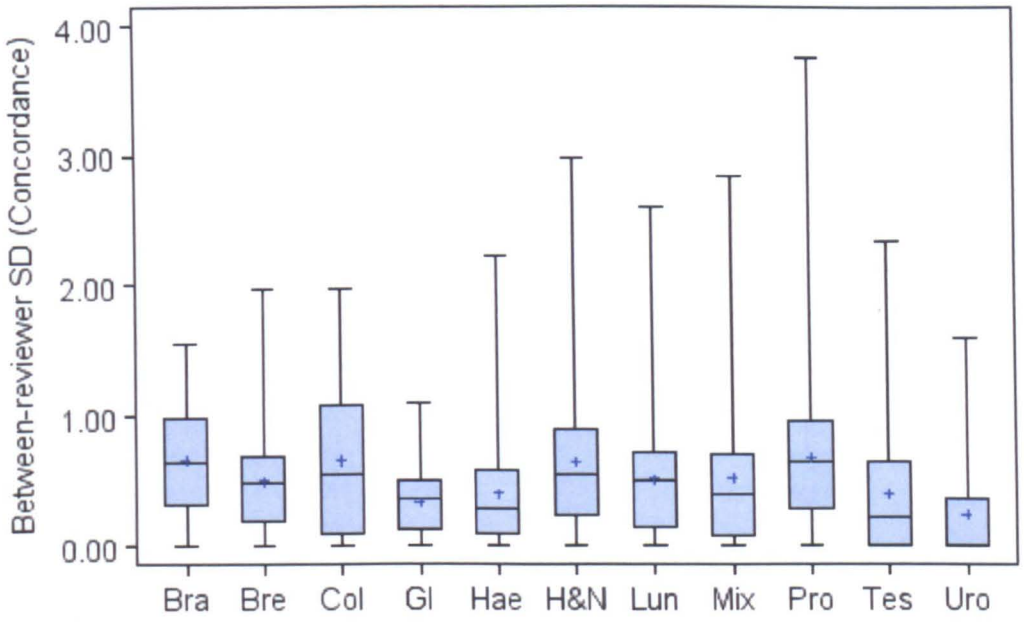


Figure 24 Between-reviewer SD by category of anchor

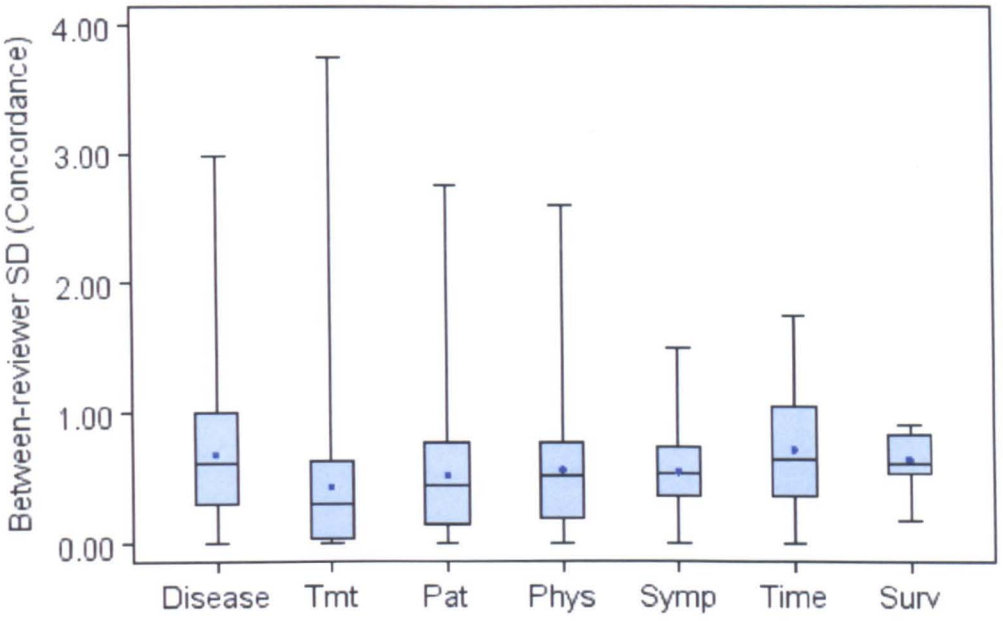


Figure 25 Between-reviewer SD by timing of contrast

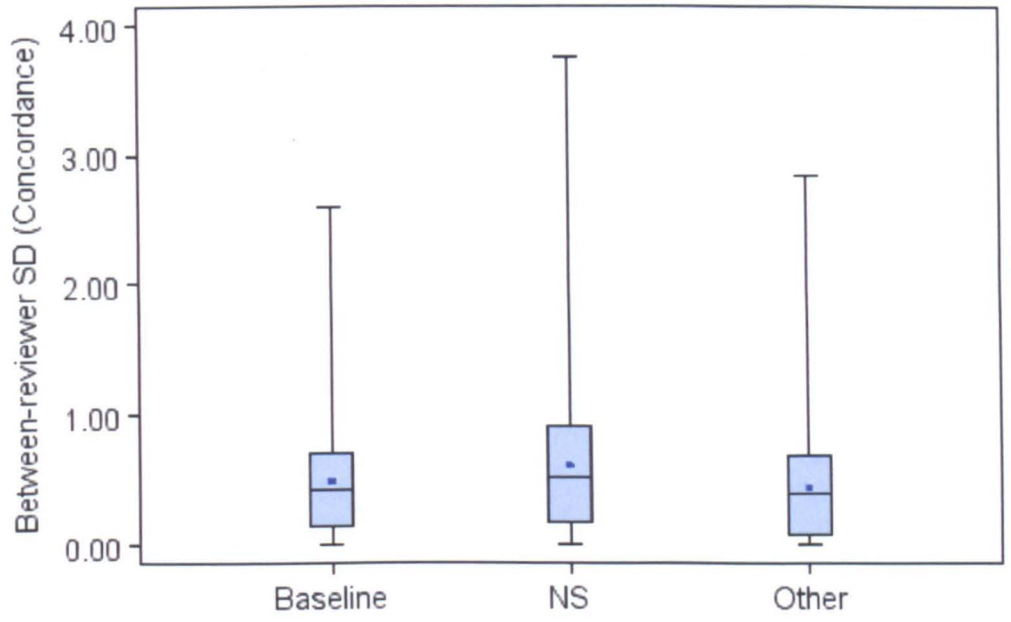


Figure 26 Between-reviewer SD by strength of anchor

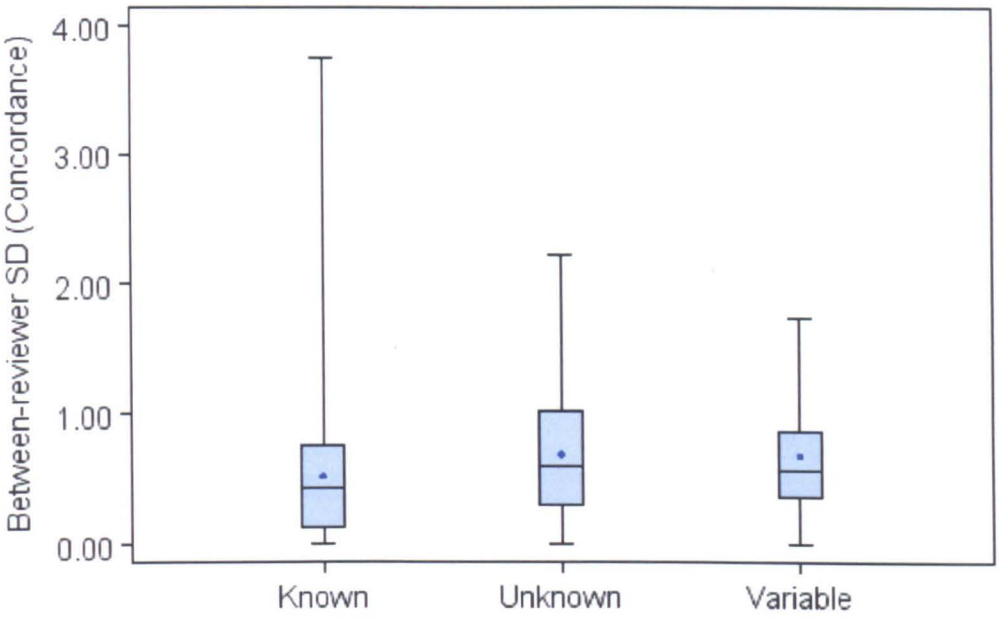


Figure 27 Between-reviewer SD by disease stage

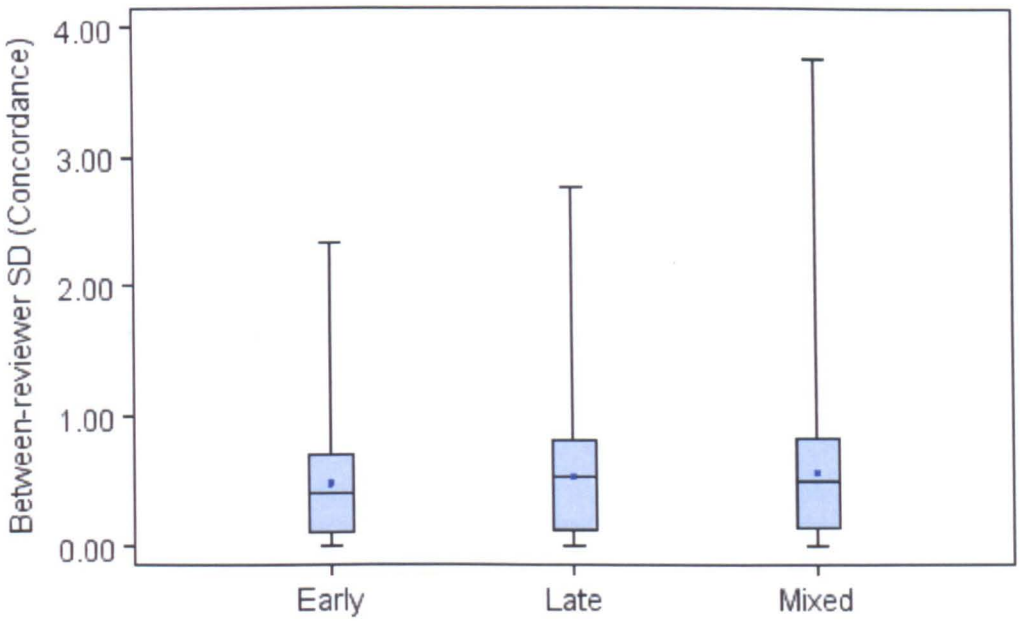


Table 45 Factors affecting concordance (cross-sectional contrasts)

Factor	Level	Estimate	SE	p-value
Study design	RCT Phase III	0.42	0.13	0.001
	Cohort	+0.17	0.02	<0.0001
	RCT Phase II	-0.15	0.12	0.21
	Multiple studies	+0.06	0.04	0.11
Cancer type	Breast	0.55	0.19	0.003
	Brain	+0.14	0.05	0.003

Factor	Level	Estimate	SE	p-value
	Colorectal	+0.15	0.03	<0.0001
	GI	+0.16	0.05	0.0009
	Haem	-0.09	0.03	0.0008
	H&N	+0.14	0.03	<0.0001
	Lung	+0.02	0.02	0.42
	Mixed	+0.01	0.03	0.62
	Prostate	+0.20	0.03	<0.0001
	Testicular	-0.07	0.05	0.16
	Urology	-0.25	0.06	<0.0001
Anchor category	Treatment	1.45	0.12	<0.0001
	Disease	+0.25	0.02	<0.0001
	Patient	+0.11	0.02	<0.0001
	Physical	+0.12	0.03	<0.0001
	Symptom	+0.11	0.04	0.002
	Time	+0.29	0.05	<0.0001
	Survival	+0.19	0.08	0.01
Disease stage	Early	0.55	0.03	<0.0001
	Late	+0.05	0.02	0.01
	Mixed	+0.11	0.02	<0.0001
Timing of contrast	Baseline	0.49	0.03	<0.0001
	Not specified	+0.14	0.02	<0.0001
	Post-baseline	-0.05	0.02	0.03
Strength of anchor	Known	0.78	0.05	<0.0001
	Variable	+0.13	0.04	0.0004
	Unknown	+0.18	0.03	<0.0001

6.2.3.2 Factors affecting concordance (longitudinal contrasts)

The factors investigated here differed slightly from those investigated for the cross-sectional contrasts. The anchor-related factors are not relevant for the contrasts over time (since the anchor here is time). Factors looking at the timing of the second time point and the dropout over time were relevant here but were not for the cross-sectional contrasts. Boxplots or scatterplots used initially to visualise how $SD_{between}$

varied across the levels of each factor can be found from Figure 28 to Figure 31. The results from the mixed models are discussed below and summarised in Table 46.

The boxplot of $SD_{between}$ by study design shows some variation in concordance across the designs. Contrasts from the cohort studies and articles containing multiple studies had significantly increased discordance compared with the RCT Phase III contrasts ($p<0.0001$). All other study designs had similar concordance to the Phase III randomised studies.

Breast cancer was used as the reference cancer type. Colorectal, haematology, H&N and testicular cancer papers had significantly higher discordance ($p<0.0001$) than breast cancer. GI and urology/kidney papers had significantly better concordance compared to breast cancer ($p<0.0001$). Contrasts from the other cancer types were similar to the breast cancer contrasts with respect to concordance.

The timing of the second time point ranged from 4 days up to a maximum of 60 months post baseline, with the majority of contrasts being within a year of the baseline. The scatterplot showing timing of the second time point against $SD_{between}$ does not show any trend for a change in concordance as the distance between time 1 and time 2 increases ($p=0.2$).

The distribution of the percentage dropout from time 1 to time 2 shows that a few contrasts have more patients at the second time than at the baseline point. These are where there were missing baseline values but subsequent data is reported. The widest range of values for $SD_{between}$ was seen where there was no dropout. The mixed model indicates that $SD_{between}$ reduces as the dropouts increase (i.e. concordance improves) which would be the opposite trend to that expected. The slope however is very shallow (a reduction in $SD_{between}$ of 0.002 per month decrease between time 1 and time 2) and the significance may simply be due to the large sample size rather than indicating a relevant change in concordance.

Figure 28 Between-reviewer SD by study design

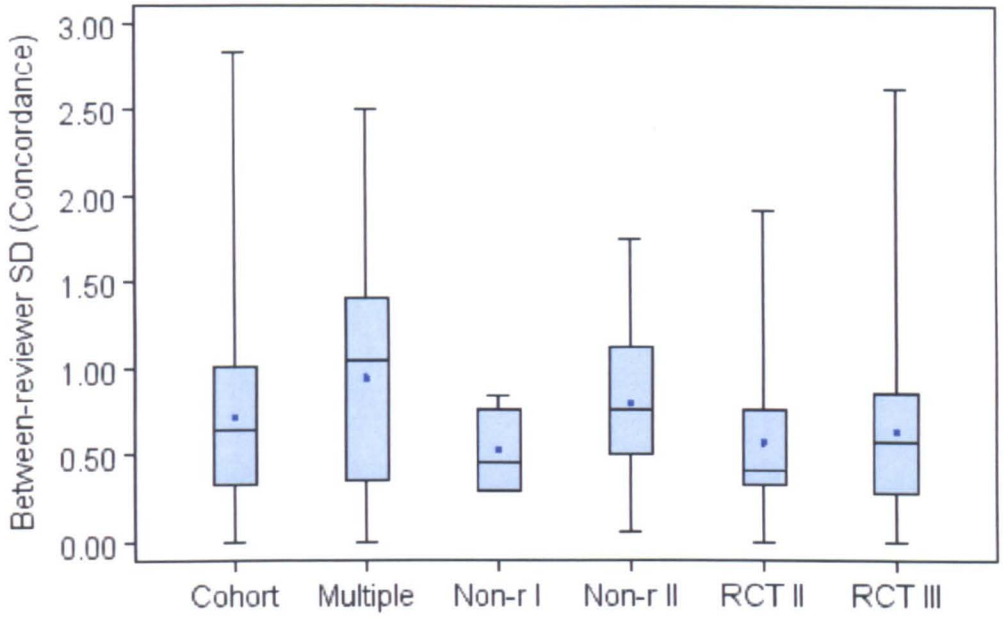


Figure 29 Between-reviewer SD by cancer type

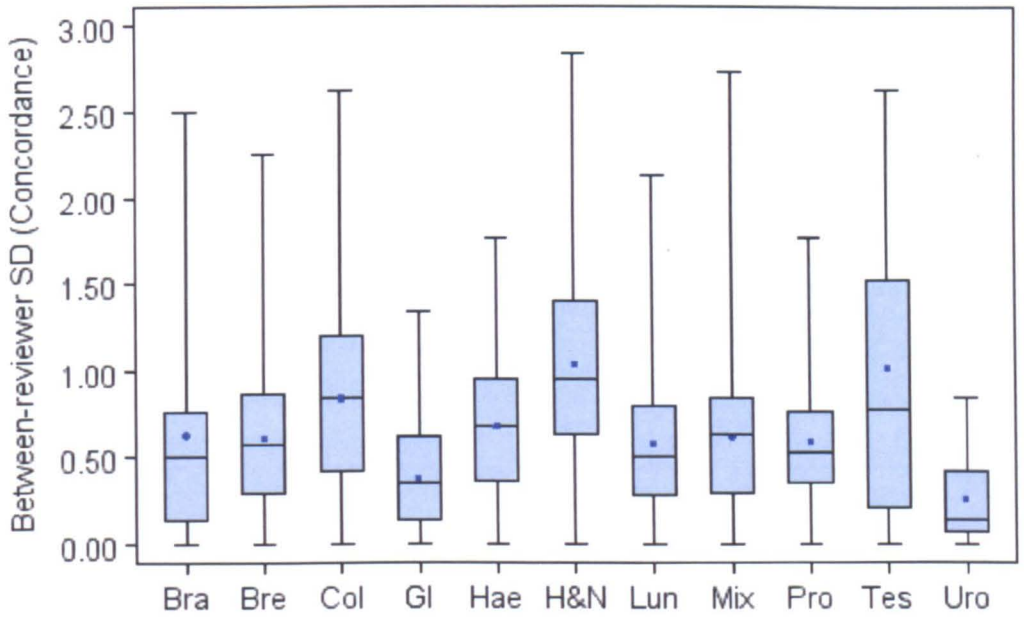


Figure 30 Between-reviewer SD by timing of second time point

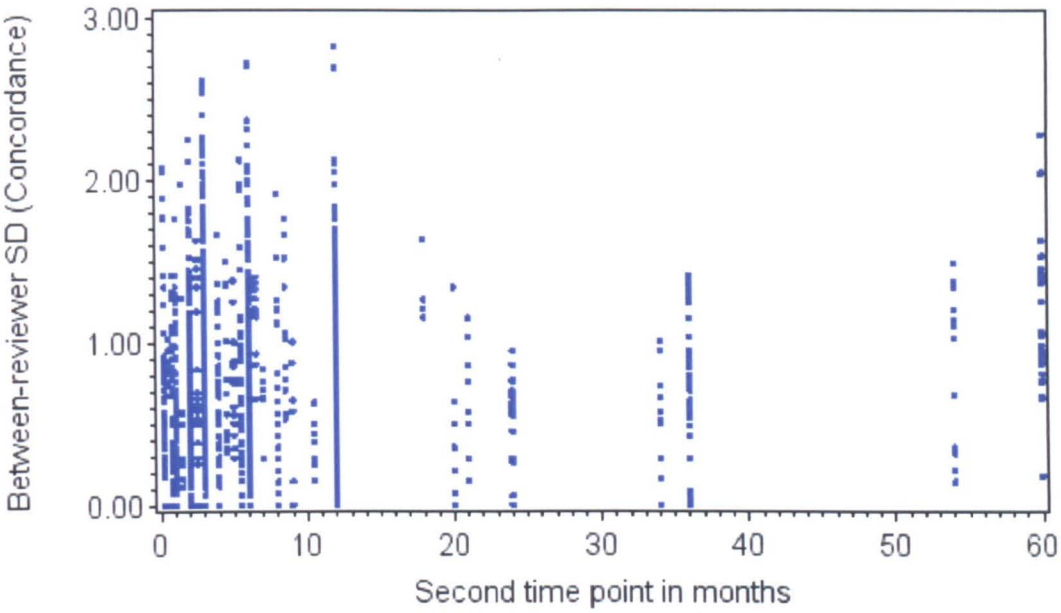


Figure 31 Between-reviewer SD by percentage dropout

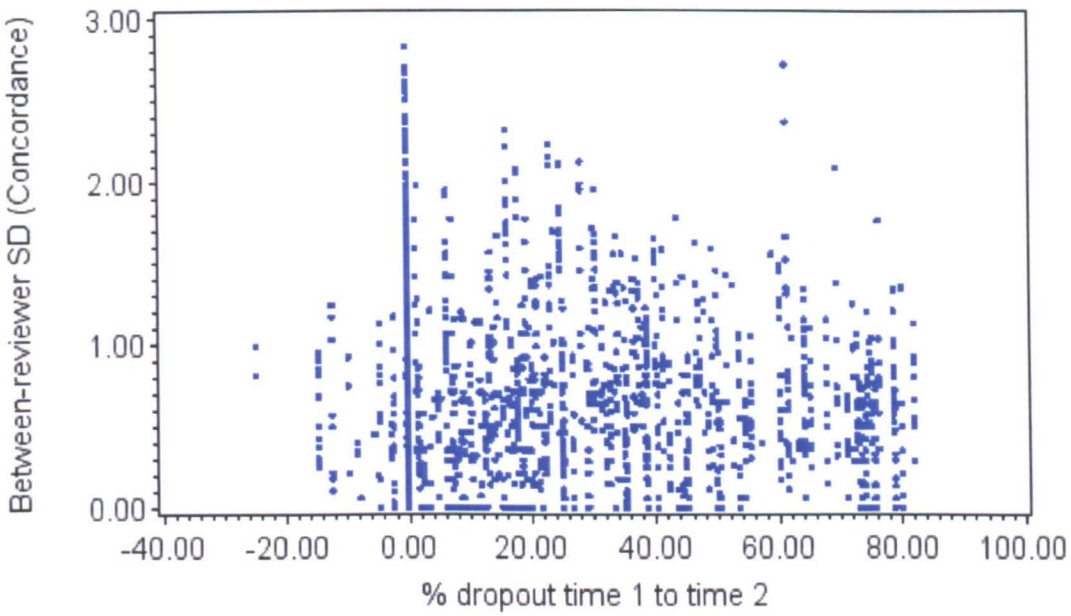


Table 46 Factors affecting concordance (longitudinal contrasts)

Factor	Level	Estimate	SE	p-value
Study design	RCT Phase III	1.01	0.15	<0.0001
	Cohort	+0.08	0.02	<0.0001
	RCT Phase II	-0.07	0.06	0.29
	Non-randomised phase I	-0.10	0.09	0.28
	Multiple studies	+0.30	0.05	<0.0001

Factor	Level	Estimate	SE	p-value
Cancer type	Breast	1.08	0.25	<0.0001
	Brain	-0.003	0.06	0.96
	Colorectal	+0.25	0.03	<0.0001
	GI	-0.21	0.05	<0.0001
	Gynae	-0.11	0.13	0.40
	Haem	+0.09	0.04	0.04
	H&N	+0.42	0.03	<0.0001
	Lung	-0.04	0.03	0.17
	Mixed	+0.02	0.03	0.58
	Prostate	-0.009	0.04	0.82
	Testicular	+0.44	0.05	<0.0001
	Urology	-0.33	0.07	<0.0001
Timing of second assessment	Intercept	0.66	0.02	<0.0001
	Slope	0.0012	0.001	0.21
% dropout	Intercept	0.67	0.02	<0.0001
	Slope	-0.002	0.0004	<0.0001

6.2.4 Uncertainty in reviewers scores

6.2.4.1 Average number of categories used by reviewers

Table 47 shows the median and range of the number of categories used (which is first averaged for each contrast across the two or three reviewers). On average the reviewers used two categories for their review of longitudinal contrasts and slightly less (1.7) for their review of the cross-sectional contrasts. There were not major differences across subscales (1.5 to 2.0). Across cancer types the haematological and lung reviewers used a higher number of categories on average. Across study designs the non-randomised phase II studies were associated with the highest average.

Table 47 Number of categories used (RCT baseline contrasts excluded)

	CONTRAST TYPE						OVERALL		
	Cross-sectional			Longitudinal					
	Median	Max	Min	Median	Max	Min	Median	Max	Min
OVERALL	1.7	4.5	1.0	2.0	5.0	1.0	2.0	5.0	1.0
SUBSCALE	1.7	4.0	1.0	2.0	4.5	1.0	1.7	4.5	1.0

	CONTRAST TYPE						OVERALL		
	Cross-sectional			Longitudinal					
	Median	Max	Min	Median	Max	Min	Median	Max	Min
AP									
CF	1.7	4.0	1.0	1.7	4.0	1.0	1.7	4.0	1.0
CO	1.5	4.0	1.0	1.7	3.5	1.0	1.7	4.0	1.0
DI	1.5	3.5	1.0	1.7	4.5	1.0	1.7	4.5	1.0
DY	1.7	4.0	1.0	2.0	4.0	1.0	1.7	4.0	1.0
EF	2.0	4.0	1.0	2.0	5.0	1.0	2.0	5.0	1.0
FA	2.0	4.0	1.0	2.0	4.0	1.0	2.0	4.0	1.0
FI	1.7	3.5	1.0	1.7	3.5	1.0	1.7	3.5	1.0
NV	1.7	4.0	1.0	2.0	4.5	1.0	1.7	4.5	1.0
PA	1.7	4.0	1.0	2.0	3.5	1.0	2.0	4.0	1.0
PF	2.0	4.0	1.0	2.0	4.5	1.0	2.0	4.5	1.0
QL	2.0	4.0	1.0	2.0	4.5	1.0	2.0	4.5	1.0
RF	2.0	4.0	1.0	2.0	5.0	1.3	2.0	5.0	1.0
SF	2.0	4.0	1.0	2.0	4.0	1.0	2.0	4.0	1.0
SL	1.7	4.5	1.0	1.7	4.0	1.0	1.7	4.5	1.0
CANCER TYPE									
Brain	1.7	3.0	1.0	2.0	3.3	1.0	2.0	3.3	1.0
Breast	1.7	3.3	1.0	1.7	3.3	1.0	1.7	3.3	1.0
Colorectal	1.3	3.0	1.0	1.7	3.5	1.0	1.5	3.5	1.0
GI	1.5	3.5	1.0	1.7	3.5	1.0	1.5	3.5	1.0
Gynaecological	.	.	.	2.0	2.5	1.0	2.0	2.5	1.0
Haematological	2.5	4.5	1.0	2.7	5.0	1.5	2.5	5.0	1.0
Head and neck	1.5	2.3	1.0	1.7	2.5	1.0	1.7	2.5	1.0
Lung	2.3	3.7	1.0	2.3	3.7	1.0	2.3	3.7	1.0
Mixed	2.0	3.0	1.0	2.0	3.5	1.0	2.0	3.5	1.0
Prostate	1.7	2.7	1.0	2.0	2.7	1.0	1.7	2.7	1.0
Testicular	1.5	3.0	1.0	1.5	3.0	1.0	1.5	3.0	1.0
Urology/Kidney	1.0	2.7	1.0	2.0	3.5	1.5	1.5	3.5	1.0
STUDY DESIGN									
Cohort	1.7	3.7	1.0	2.0	5.0	1.0	2.0	5.0	1.0
Multiple	2.0	3.0	1.0	1.5	3.0	1.3	1.7	3.0	1.0

	CONTRAST TYPE						OVERALL		
	Cross-sectional			Longitudinal					
	Median	Max	Min	Median	Max	Min	Median	Max	Min
Non-randomised Phase I	.	.	.	1.7	1.7	1.3	1.7	1.7	1.3
Non-randomised Phase II	.	.	.	2.7	3.0	1.7	2.7	3.0	1.6
RCT II	1.5	2.0	1.0	1.8	3.5	1.0	1.7	3.5	1.0
RCT III	2.0	4.5	1.0	2.0	4.0	1.0	2.0	4.5	1.0

6.2.4.2 Peak weighting used by reviewers

Table 48 shows the median and range of the peak weighting used (which is first averaged for each contrast across the two or three reviewers). The overall average is high (80%), with longitudinal contrasts having a slightly lower average peak (75%) compared with the cross-sectional contrasts (80%). Across subscales the average peak ranged from 72.5% to 90.0%. The subscale with the highest average peak was DI and the lowest average was PF. The average peak varied more across cancer types, with haematological contrasts having the lowest average of 60.0% compared to the highest average of 100% for the testicular contrasts. Across study designs the average ranged from 57.5% for non-randomised phase II studies up to 90.0% for the randomised phase II studies.

Table 48 Peak weighting (RCT baseline contrasts excluded)

	CONTRAST TYPE						OVERALL		
	Cross-sectional			Longitudinal					
	Median	Max	Min	Median	Max	Min	Median	Max	Min
OVERALL	80.0	100.0	30.0	75.0	100.0	30.0	76.7	100.0	30.0
SUBSCALE									
AP	83.3	100.0	37.5	76.7	100.0	42.0	80.0	100.0	37.5
CF	83.3	100.0	35.0	83.3	100.0	40.0	83.3	100.0	35.0
CO	86.7	100.0	35.0	83.3	100.0	50.0	85.8	100.0	35.0
DI	90.0	100.0	35.0	83.3	100.0	45.0	86.7	100.0	35.0
DY	83.3	100.0	35.0	80.0	100.0	43.3	83.3	100.0	35.0
EF	76.7	100.0	35.0	75.0	100.0	30.0	75.0	100.0	30.0
FA	75.0	100.0	35.0	73.3	100.0	40.0	75.0	100.0	35.0

	CONTRAST TYPE								
	Cross-sectional			Longitudinal			OVERALL		
	Median	Max	Min	Median	Max	Min	Median	Max	Min
FI	85.0	100.0	45.0	85.0	100.0	40.0	85.0	100.0	40.0
NV	83.3	100.0	37.5	80.0	100.0	50.0	80.8	100.0	37.5
PA	80.0	100.0	35.0	75.0	100.0	46.7	76.7	100.0	35.0
PF	75.0	100.0	40.0	72.5	100.0	40.0	75.0	100.0	40.0
QL	75.0	100.0	40.0	73.3	100.0	34.5	75.0	100.0	34.5
RF	75.0	100.0	40.0	75.0	97.5	35.0	75.0	100.0	35.0
SF	76.7	100.0	37.5	75.0	100.0	40.0	75.0	100.0	37.5
SL	83.3	100.0	30.0	80.0	100.0	37.5	83.3	100.0	30.0
CANCER TYPE									
Brain	76.7	100.0	58.3	83.3	100.0	53.3	80.0	100.0	53.3
Breast	80.0	100.0	48.3	76.7	100.0	50.0	76.7	100.0	48.3
Colorectal	90.0	100.0	60.0	83.3	100.0	45.0	85.0	100.0	45.0
GI	80.0	100.0	37.5	75.0	100.0	34.5	75.0	100.0	34.5
Gynaecological	.	.	.	75.0	100.0	75.0	75.0	100.0	75.0
Haematological	70.0	100.0	30.0	60.0	97.5	30.0	67.5	100.0	30.0
Head and neck	83.3	100.0	56.7	83.3	100.0	50.0	83.3	100.0	50.0
Lung	70.0	100.0	43.3	66.7	100.0	40.0	70.0	100.0	40.0
Mixed	78.8	100.0	45.0	75.0	100.0	57.5	75.8	100.0	45.0
Prostate	80.0	100.0	50.0	75.0	100.0	50.0	79.2	100.0	50.0
Testicular	82.5	100.0	50.0	95.0	100.0	50.0	90.0	100.0	50.0
Urology/Kidney	100.0	100.0	61.7	75.0	95.0	45.0	87.5	100.0	45.0
STUDY DESIGN									
Cohort	80.0	100.0	43.3	75.0	100.0	30.0	76.7	100.0	30.0
Multiple	75.0	100.0	55.0	75.0	93.3	55.0	75.0	100.0	55.0
Non-randomised Phase I	.	.	.	80.0	83.3	66.7	80.0	83.3	66.7
Non-randomised Phase II	.	.	.	57.5	83.3	45.0	57.5	83.3	45.0
RCT II	90.0	100.0	75.0	71.7	100.0	34.5	75.0	100.0	34.5
RCT III	80.0	100.0	30.0	80.0	100.0	35.0	80.0	100.0	30.0

6.2.4.3 Within-reviewer SD

The average uncertainty (as measured by the within-reviewer SD) was 0.41 for the cross-sectional and 0.46 for the longitudinal contrasts. Uncertainty was consistently worse across all subscales for the longitudinal contrasts although some of the differences were relatively small.

Table 49 Within-reviewer standard deviation (uncertainty)

	Cross-sectional							Longitudinal						
	Mean	Median	SD	SE	Max	Min	N	Mean	Median	SD	SE	Max	Min	N
Subscale														
AP	0.37	0.37	0.24	0.02	1.04	0.00	249	0.45	0.43	0.23	0.02	1.13	0.00	199
CF	0.38	0.35	0.24	0.01	1.17	0.00	294	0.39	0.37	0.19	0.01	0.99	0.00	231
CO	0.32	0.29	0.26	0.02	1.11	0.00	228	0.32	0.32	0.23	0.02	0.90	0.00	175
DI	0.26	0.24	0.22	0.02	1.04	0.00	217	0.33	0.29	0.24	0.02	1.09	0.00	176
DY	0.35	0.32	0.25	0.02	1.17	0.00	213	0.43	0.41	0.25	0.02	0.95	0.00	177
EF	0.45	0.44	0.24	0.01	1.15	0.00	344	0.47	0.45	0.21	0.01	1.41	0.00	267
FA	0.47	0.46	0.23	0.01	1.15	0.00	292	0.53	0.49	0.20	0.01	1.06	0.00	268
FI	0.32	0.29	0.22	0.02	1.04	0.00	196	0.36	0.32	0.21	0.02	1.06	0.00	130
NV	0.37	0.35	0.24	0.02	1.04	0.00	248	0.43	0.43	0.21	0.01	0.98	0.00	215
PA	0.43	0.41	0.24	0.01	1.18	0.00	288	0.46	0.45	0.21	0.01	0.95	0.00	230
PF	0.49	0.48	0.22	0.01	1.07	0.00	336	0.52	0.53	0.21	0.01	1.07	0.00	284
QL	0.48	0.49	0.22	0.01	1.09	0.00	390	0.54	0.53	0.22	0.01	1.17	0.00	304
RF	0.48	0.47	0.22	0.01	1.02	0.00	328	0.49	0.46	0.19	0.01	1.28	0.14	263
SF	0.45	0.45	0.23	0.01	1.04	0.00	354	0.50	0.47	0.19	0.01	1.14	0.00	279
SL	0.38	0.35	0.23	0.02	1.20	0.00	238	0.41	0.40	0.21	0.02	1.11	0.00	199
All	0.41	0.41	0.24	0.00	1.20	0.00	4215	0.46	0.44	0.22	0.00	1.41	0.00	3397

6.2.5 Factors affecting uncertainty

The SD_{within} as defined in section 4.4.3.4.3 was used as the outcome variable to investigate which factors may be associated with greater uncertainty. Note that a high SD_{within} indicates more uncertainty and a low SD_{within} indicates more certainty.

Boxplots were used initially to visualise how SD_{within} varied across the categories of each factor. The boxplots use a '+' symbol to indicate the mean and a line across the box for the median value. The box represents the inter-quartile range and the whiskers represent the minimum and maximum values.

Tables were used to show the results from mixed models used to investigate the significance of each factor. In the table, an estimate with a '+' sign this indicates a higher SD_{within} (more uncertainty) compared with the reference category for that factor, whereas a '-' sign indicates a lower SD_{within} or greater certainty.

6.2.5.1 Factors affecting uncertainty (cross-sectional contrasts)

Figure 32 to Figure 37 show how uncertainty (as measured by the within-reviewer SD) varied across levels of the factors. Table 50 summarises the results from the mixed models. There were some levels in each of the factors with significantly different levels of uncertainty to the chosen reference levels.

In terms of study design only the cohort studies differed significantly from the RCT Phase III studies ($p<0.0001$), with more uncertainty for the contrasts from cohort studies.

The boxplot shows there was considerable variation in uncertainty across cancer types. Haematology, prostate and lung cancer contrasts had significantly more uncertainty than breast cancer ($p<0.0001$). Colorectal, H&N, testicular and urology/kidney contrasts had significantly less uncertainty than the breast cancer contrasts ($p<0.0001$).

For the anchor types, all categories had higher uncertainty than the treatment-related anchors. This increase was significant for all except the survival category ($p<0.0001$ for disease, patient, physical, symptom and time-related anchors).

The contrasts from patients with late stage disease had significantly less uncertainty than those from early stage disease (<0.0001). Contrasts from early stage patients and a mixture of early and late stage patients were similar in terms of uncertainty.

Baseline contrasts had the lowest uncertainty. Contrasts at other points in time and those where the timing was not clear from the full article had higher uncertainty ($p<0.0001$).

Anchors with an unknown link to QOL had significantly higher uncertainty than the known anchors ($p=0.0007$). However, the anchors we classed as 'variable' had significantly less uncertainty than the known anchors ($p=0.0003$).

Figure 32 Within-reviewer SD by study design

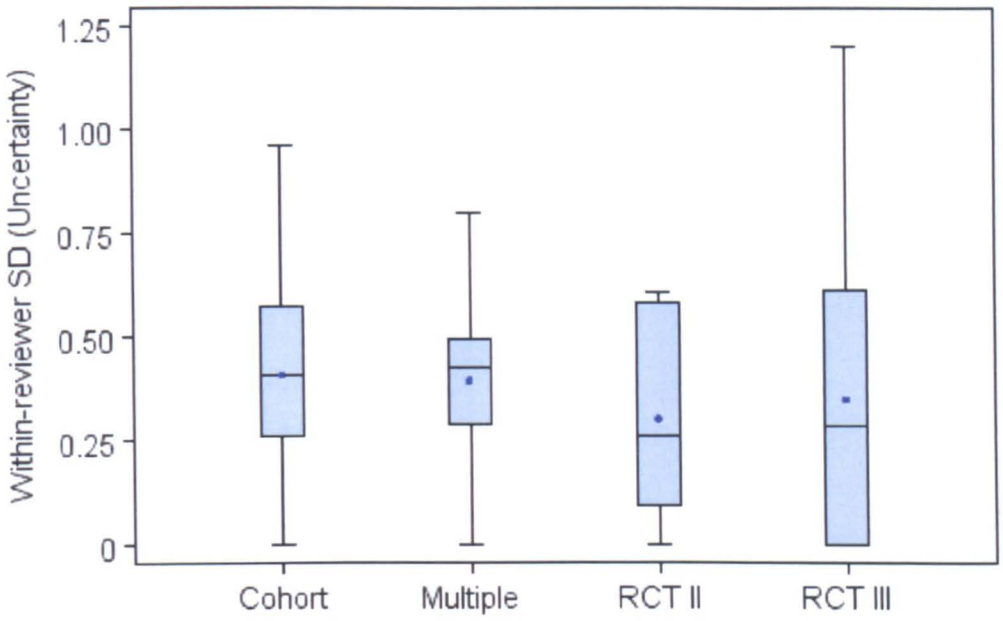


Figure 33 Within-reviewer SD by cancer type

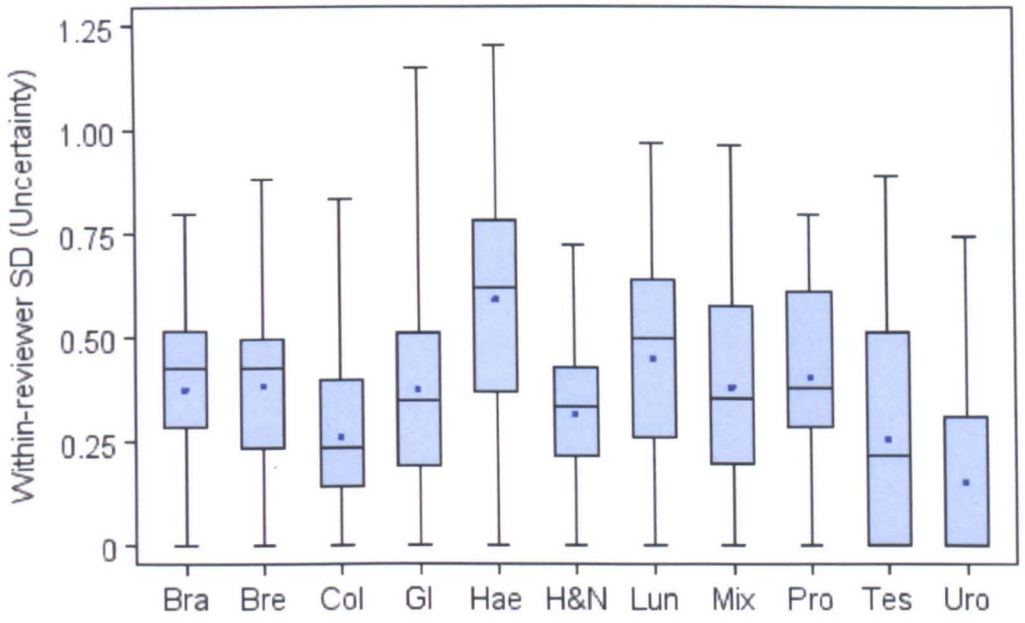


Figure 34 Within-reviewer SD by category of anchor

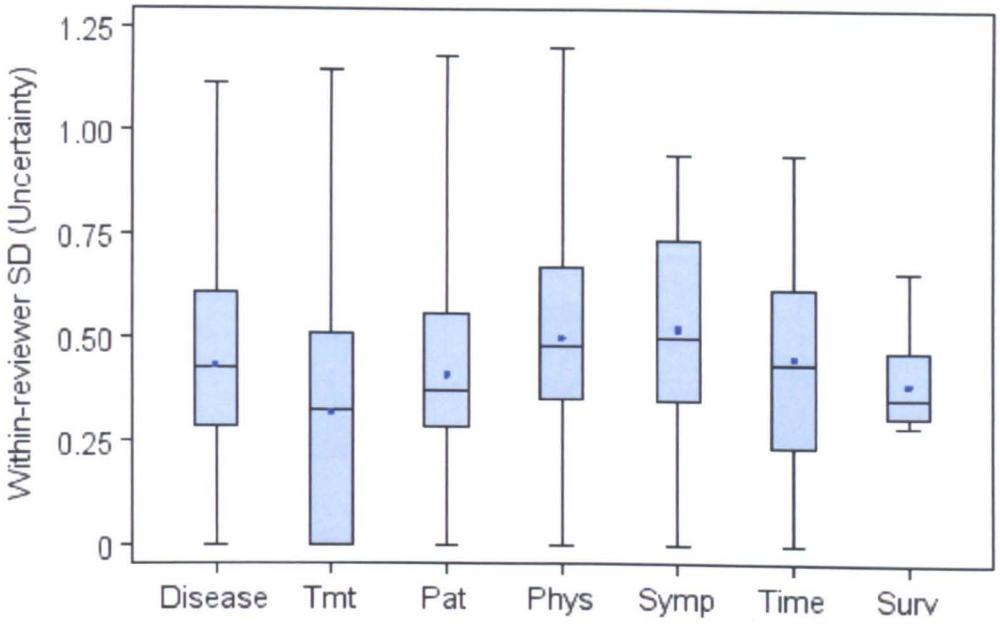


Figure 35 Within-reviewer SD by timing of contrast

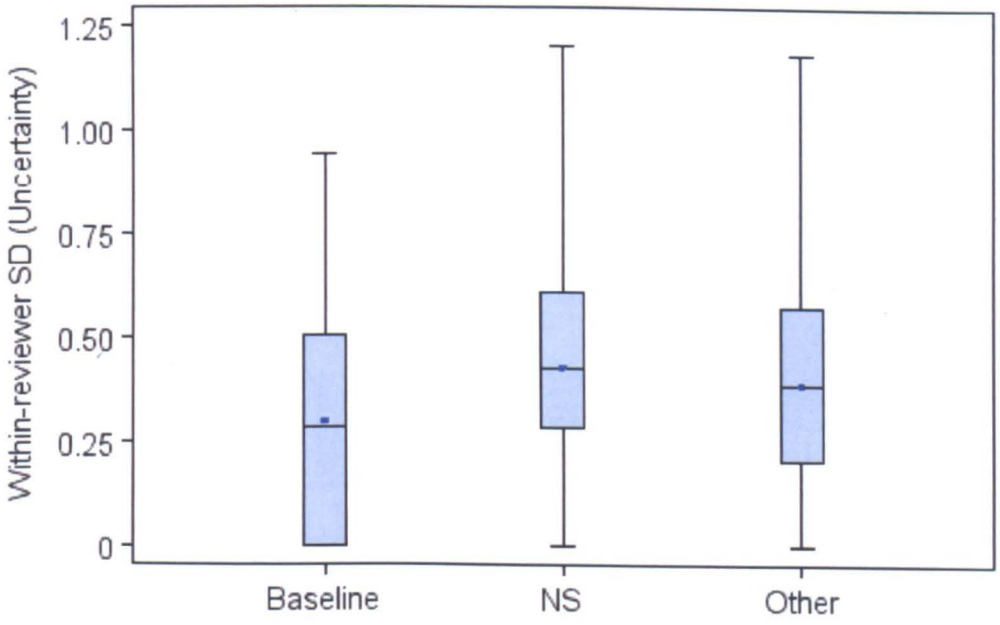


Figure 36 Within-reviewer SD by strength of anchor

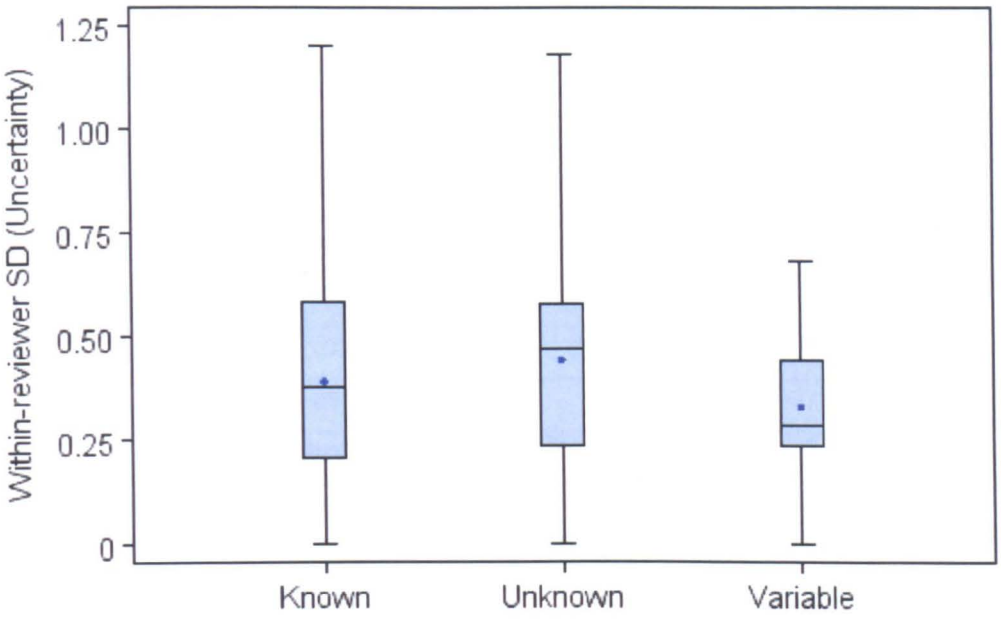


Figure 37 Within-reviewer SD by disease stage

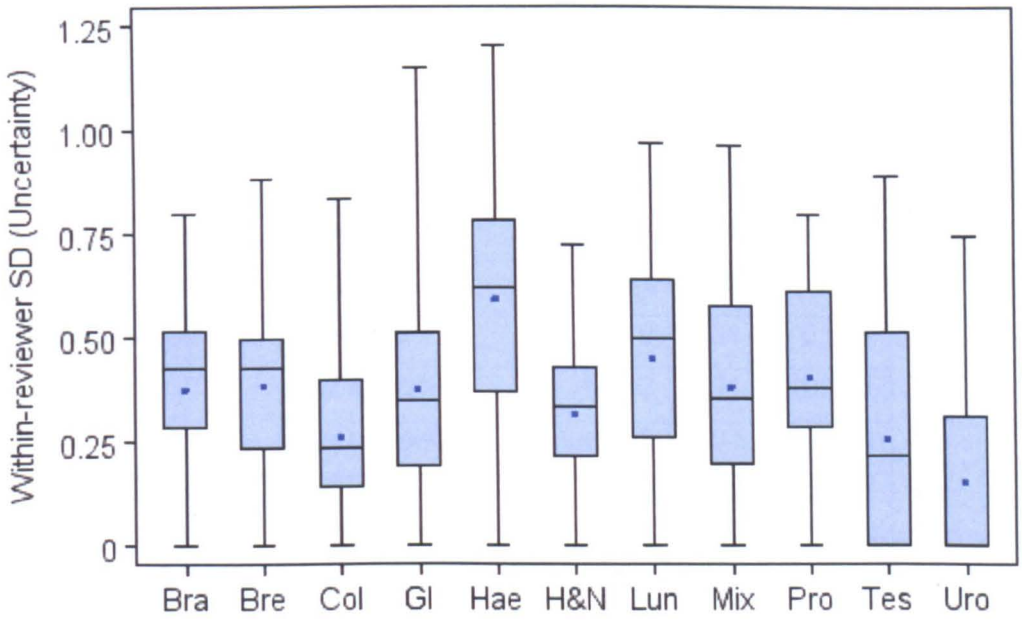


Table 50 Factors affecting uncertainty (cross-sectional contrasts)

Factor	Level	Estimate	SE	p-value
Study design	RCT Phase III	0.36	0.07	<0.001
	Cohort	+0.06	0.01	<0.001
	RCT Phase II	-0.06	0.06	0.33
	Multiple studies	+0.04	0.02	0.09
Cancer type	Breast	0.07	0.09	0.46
	Brain	-0.04	0.02	0.12

Factor	Level	Estimate	SE	p-value
	Colorectal	-0.13	0.02	<0.001
	GI	-0.01	0.02	0.80
	Haem	+0.21	0.01	<0.001
	H&N	-0.09	0.01	<0.001
	Lung	+0.07	0.01	<0.001
	Mixed	-0.01	0.01	0.45
	Prostate	+0.04	0.01	0.006
	Testicular	-0.09	0.02	<0.001
	Urology	-0.23	0.03	<0.001
Anchor category	Treatment	1.11	0.06	<0.001
	Disease	+0.11	0.01	<0.001
	Patient	+0.09	0.01	<0.001
	Physical	+0.18	0.01	<0.001
	Symptom	+0.20	0.02	<0.001
	Time	+0.13	0.03	<0.001
	Survival	+0.07	0.04	0.07
Disease stage	Early	0.30	0.01	<0.001
	Late	-0.08	0.01	<0.001
	Mixed	-0.001	0.01	0.87
Timing of contrast	Baseline	0.49	0.01	<0.001
	Not specified	+0.13	0.01	<0.001
	Post-baseline	+0.08	0.01	<0.001
Strength of anchor	Known	0.34	0.03	<0.001
	Variable	-0.07	0.02	0.003
	Unknown	+0.06	0.02	0.007

6.2.5.2 Factors affecting uncertainty (longitudinal contrasts)

Figure 38 to Figure 41 show how uncertainty (as measured by the within-reviewer SD) varied across levels of the factors. Table 51 summarises the results from the mixed models. The boxplots show considerable variation in uncertainty for study designs and cancer types. The scatterplots do not show any clear trends for an increase or decrease in uncertainty as the second time point gets later or the dropout increases.

Contrasts from cohort studies and from phase II studies (both randomised and non-randomised) had significantly higher uncertainty than the Phase III RCTs ($p<0.0001$ to $p=0.002$). The other study designs did not differ significantly with respect to uncertainty from the Phase III RCTs.

Colorectal, H&N and testicular cancer sites had less uncertainty than the breast cancer contrasts ($p<0.0001$). Haematology, lung, prostate and urology/kidney had significantly more uncertainty ($p<0.0001$ to 0.0005).

The results from the mixed models indicated that there was a decrease in uncertainty as the timing of the second time point increased from baseline ($p<0.0001$) and as the percentage dropout increased ($p<0.0001$). However, the slope in both models is very shallow and the statistical significance may be a result of the high numbers of contrasts rather than indicating relevant changes in uncertainty.

Figure 38 Within-reviewer SD by study design

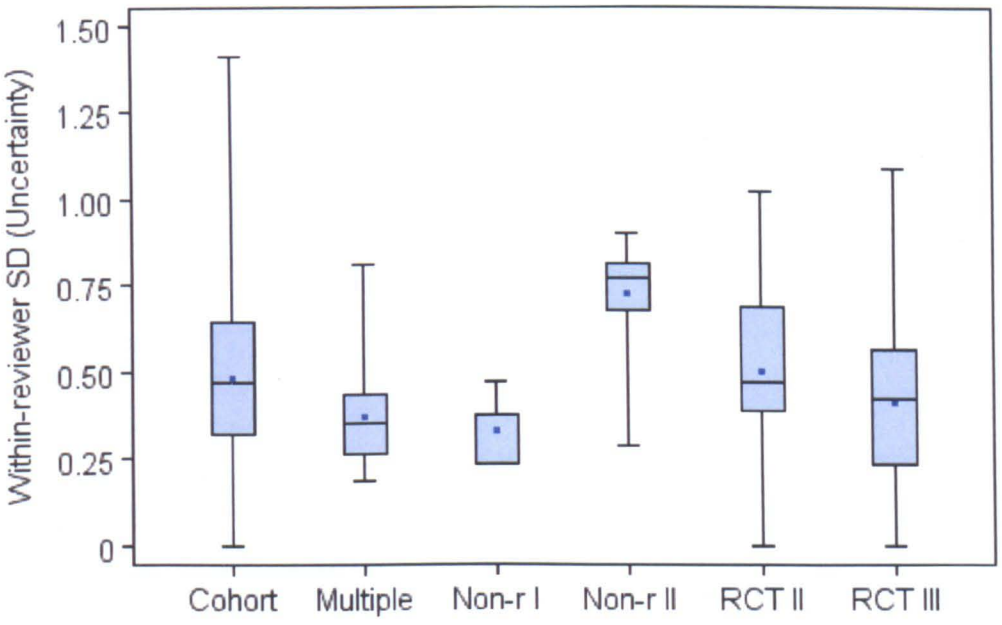


Figure 39 Within-reviewer SD by cancer type

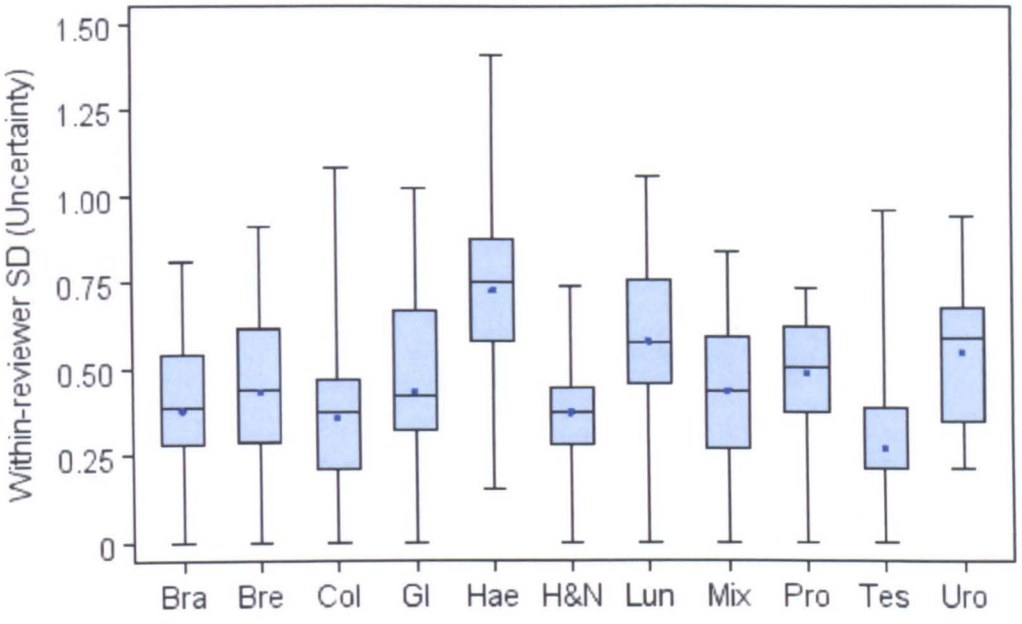


Figure 40 Within-reviewer SD by timing of second time point

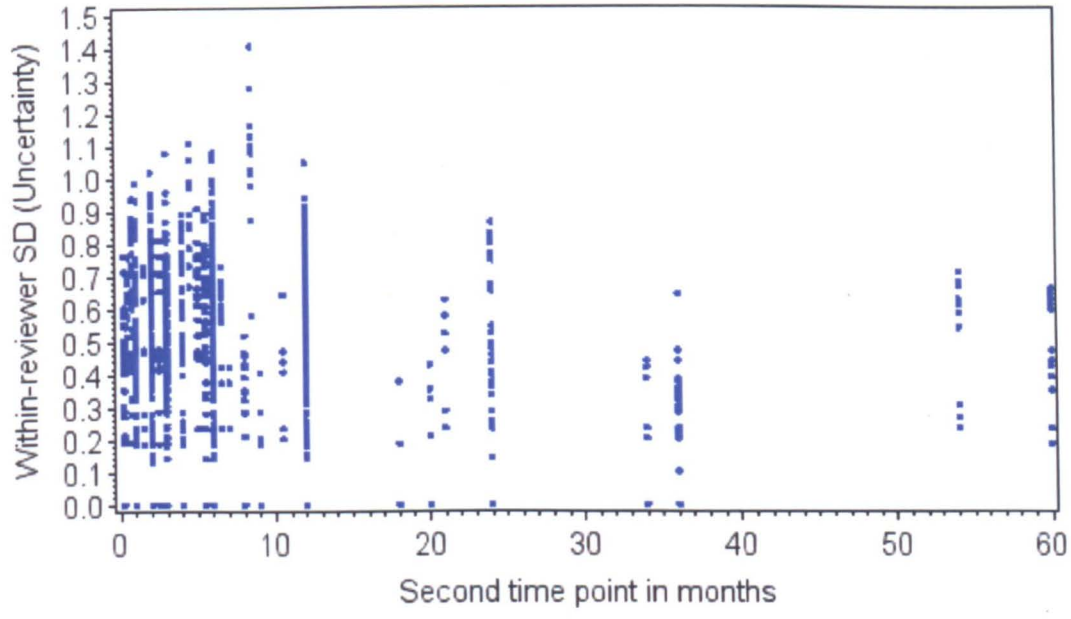


Figure 41 Within-reviewer SD by percentage dropout

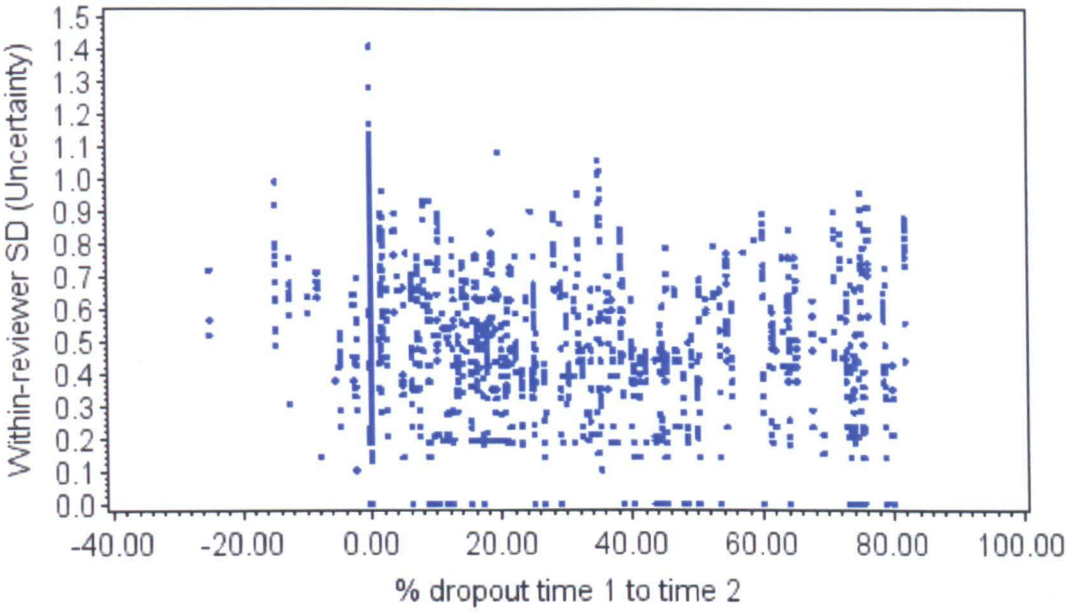


Table 51 Factors affecting uncertainty (longitudinal contrasts)

Factor	Level	Estimate	SE	p-value
Study design	RCT Phase III	0.75	0.07	<0.001
	Cohort	+0.07	0.01	<0.001
	RCT Phase II	+0.09	0.03	0.002
	Non-randomised phase I	-0.08	0.04	0.05
	Non-randomised phase II	+0.31	0.03	<0.001
	Multiple studies	-0.05	0.02	0.02
Cancer type	Breast	0.55	0.10	<0.001
	Brain	-0.06	0.02	0.02
	Colorectal	-0.08	0.01	<0.001
	GI	-0.001	0.02	0.95
	Gynae	-0.13	0.05	0.02
	Haem	+0.29	0.02	<0.001
	H&N	-0.06	0.01	<0.001
	Lung	+0.14	0.01	<0.001
	Mixed	0.00	0.01	0.94
	Prostate	+0.05	0.02	0.005
	Testicular	-0.16	0.02	<0.001

Factor	Level	Estimate	SE	p-value
	Urology	+0.11	0.03	<0.001
Timing of second assessment	Intercept	0.45	0.01	<0.001
	Slope	-0.003	0.0004	<0.001
% dropout	Intercept	0.46	0.007	<0.001
	Slope	-0.001	0.0002	<0.001

6.2.6 Summary and Conclusions

The cross-sectional contrasts had slightly better correlation with the scores from the papers then the longitudinal contrasts. When considering all contrasts together the physical functioning and fatigue subscales had the highest correlation between the overall opinion from the experts and the actual QOL scores. The financial impact subscale had almost zero correlation between the experts and the actual QOL scores therefore it is questionable as to whether guidelines for this subscale can realistically be developed using this method.

Some individual reviewers had close to zero or a negative correlation with the actual QOL differences. However as these reviewers only reviewed a small number of papers the impact on the analysis should be minimal.

Agreement between reviewers on the same contrast was fairly high, with 76% of cross-sectional and 63% of longitudinal contrasts with the maximum distance between reviewers of one size class or less. There were less than 10% of contrasts with reviewers a distance of more than three categories apart.

Generally the certainty attached to the reviews was high (with an average of two categories used and an average peak weight of 80%).

All of the factors investigated had some influence on concordance and uncertainty. Across cross-sectional and longitudinal contrasts only study design and cancer type were the factors relevant to both. In terms of study design, contrasts from cohort studies had the worst concordance between reviewers and the highest uncertainty. For cancer types, urology/kidney consistently had better concordance and uncertainty when compared with breast cancer. The cancer-type factor is confounded with reviewer (since there was a subset of reviewers reviewing in each disease area). Therefore the observation that cancer site influences both concordance and uncertainty may be due to the papers within disease areas, the reviewers who took part in reviewing them or a combination of the two. For urology/kidney there were only three reviewers so all papers will have been reviewed by the same reviewers and we have

shown that they had consistently higher concordance and certainty. Similarly for brain cancer there were only three reviewers and two of these had poor correlation with the actual QOL scores. Brain cancer was shown to have among the worst concordance between the reviewers compared with the other disease types.

Additionally for the cross-sectional contrasts non-treatment related anchors had poorer concordance and higher uncertainty, as did contrasts where the timing was unclear from the full article and anchors not considered well-known (i.e. strongly associated with QOL). For the longitudinal contrasts there were no additional factors that consistently affected both concordance and uncertainty.

There was an indication that concordance may be better when the level of dropout is higher. Initially one might expect that judging the expected QOL difference may be harder in the presence of dropout. However, it is well known that dropout in QOL studies is informative, i.e. the patients dropping out are more likely to be the sicker patients with poorer QOL. It may have been the case that where attrition was higher the experts knew that the subset remaining would have better QOL than those dropping out and therefore may actually find the judgments easier thus leading to improved concordance between the reviewers. This highlights the importance of including the contrasts with high attrition, whereas the EBES project excluded studies with high attrition from the project. It may be that these provide important contrasts to contribute to the guidelines.

6.2.7 Exclusion of poor quality contrasts

As summarised above, the agreement between reviewers and the certainty with which the reviewers judged the QOL differences seems high, indicating quality reviews for the meta-analysis. The main issue is with correlation between expert scores and the score from the original article and this led to a decision to exclude some of the poorer quality contrasts from the analysis.

The scatterplots of mean difference from the original article against the average expert score showed that there were a subset of contrasts where the actual mean differences between groups was in one direction (e.g. group A better than group B) and the expert opinion was in the opposite direction. We could speculate as to why this may occur; confusion in assessing direction by the experts, quality of the study the contrast belongs to, chance results from the study, small sample size, anchor not closely related to QOL, insufficient detail in the original article for experts to base their opinion on could all contribute. However, regardless of the reason, grouping contrasts using the

expert opinion places the mean difference from these contrasts in a completely inappropriate category and therefore we felt the resulting estimates from the meta-analysis would be unreliable. For example, say a longitudinal contrast had a mean difference of 10 points which represented an improvement for patients but the expert size class grouped this with other mean differences representing medium deteriorations over time. Given a number of these contrasts in the medium deterioration size class the estimate from the meta-analysis would be brought closer to zero.

In order to get the best possible estimates from the meta-analysis it was clear that contrasts with this particular issue should be excluded. The dataset may still contain other data where the relationship between average expert score and actual mean difference was possibly incorrect (such as contrasts with a medium expert size class and actually a zero mean difference for example). However, there is no way to exclude contrasts of this nature without making an assumption about the very size of differences we were trying to estimate. Assuming there are errors of this nature at both ends of the range (i.e. very low scores and very high scores placed in the size class) then the use of the mean estimate should be able to allow for this.

I had originally planned a sensitivity analysis excluding the reviewers with negative or close to zero correlation. However, by excluding the poor quality contrasts defined in this way those specific reviewers no longer contributed to the analysis anyway.

This exclusion criterion only applies to the contrasts placed in the small, medium and large categories by the expert review. The contrasts placed in the trivial size class do not have the same concept of direction (in theory they should be differences that have some variation around zero) therefore one cannot exclude any based on discrepancies in direction between the papers and experts. The trivial differences therefore had to be treated in a slightly different way when it came to assessing their quality. However, the exclusion criteria had the same aim of excluding only contrasts where there was clearly something wrong with either the score in the paper or with the expert size class. I thought that the main reason for 'error' in this size class would be where contrasts with a likely real difference were placed in the trivial category because of averaging across experts. For example, a contrast with two reviewers and individual weighted averages of -2 and 2 would be placed in the trivial group as the average of the two reviews is zero. However, in reality, this probably represents a contrast with a medium sized difference but the reviews showed uncertainty as to which way the

difference would be. It was impossible to identify the contrasts with this problem because, as we found earlier, it would require making an assumption about a cut-off for a meaningful difference. For example, possible scenarios include weighted averages at the opposite extremes such as -2 and 2 but also reviews such as 0.5 and -0.1. In order to exclude reviews where the direction was unclear we would have to decide which were clear errors and which were just variation around the mean. Therefore the exclusion criteria for contrasts in the trivial size class was based on obtaining the contrasts with the best agreement between reviewers rather than on the direction of the scores. This was decided since the completely opposing reviews would have high levels of disagreement. Only contrasts where the weighted averages from each individual expert placed the contrast in the trivial size class were therefore included in the analysis dataset for the trivial size class. Although this is quite harsh as a criterion and some genuinely trivial differences are probably being excluded, it avoided introducing further subjectivity by having to define a level that represented poor quality on one of the agreement measures.

6.2.8 Correlation, concordance and uncertainty in final analysis dataset

Concordance and uncertainty were not altered much by applying the agreement criteria (full details not shown). There were 75% of the cross-sectional contrasts with a maximum distance of one category between them and 67% of the longitudinal contrasts. This is similar to the full dataset for the cross-sectional contrasts and improved slightly in the analysis dataset for the longitudinal contrasts. The median number of categories used for reviews was two (as it was for the full dataset). The average peak weighting was also similar (77.5% overall).

The correlation statistics were improved as we would expect since we had removed the observations contributing to poor correlation. The graphs of expert average score versus mean difference are shown however for comparison with the full dataset shown earlier. The trivial differences are now only those where all experts agree (hence the average score is 0 and expert average scores between zero and 0.5 no longer exist). There is still a considerable amount of scatter around the rest of the scale but removing the contrasts where experts and the paper disagreed on direction means that there is a more positive relationship between the two for both cross-sectional (Figure 42) and longitudinal contrasts (Figure 43).

Figure 42 Expert average scores versus mean difference from papers (cross-sectional analysis dataset)

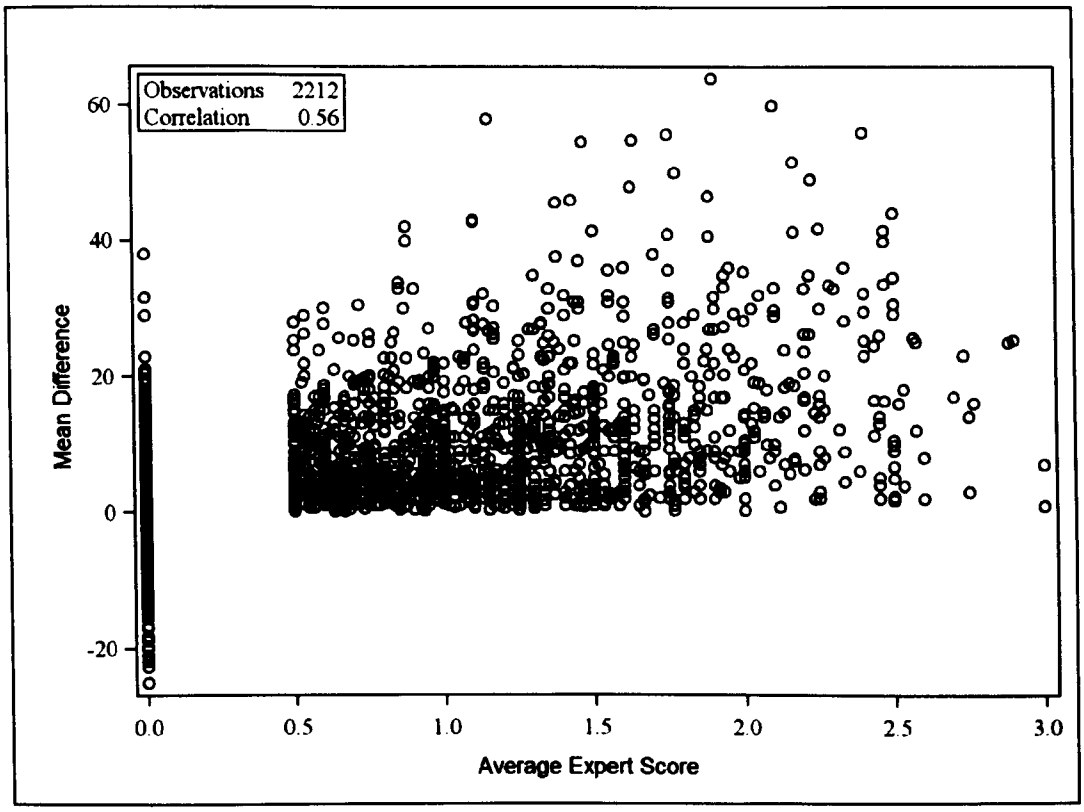
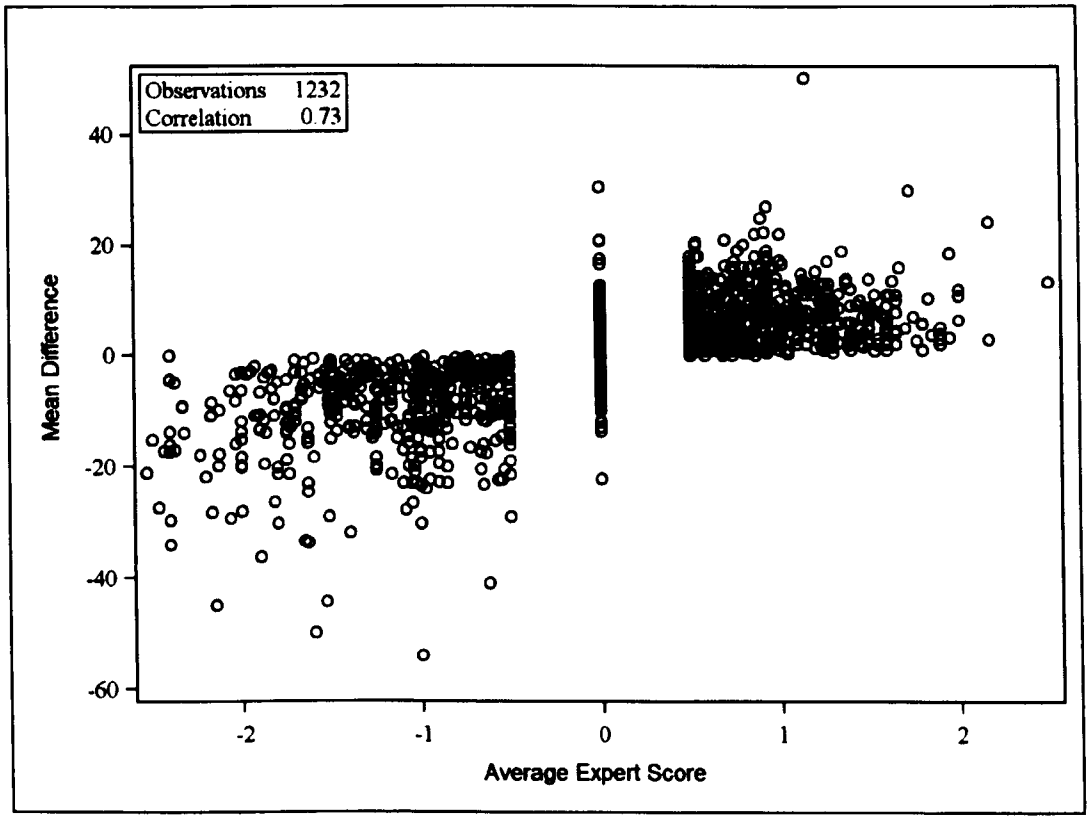


Figure 43 Expert average scores versus mean difference from papers (longitudinal analysis dataset)



7 Meta-analysis results and Evidence-Based Interpretation Guidelines

The meta-analysis results and guidelines for the cross-sectional contrasts have been published in the Journal of Clinical Oncology (2011), with full reference as follows:-

Kim Cocks, Madeleine T. King, Galina Velikova, Marrissa Martyn-St-James, Peter M. Fayers and Julia M. Brown. "Evidence-Based Guidelines for the Determination of Sample Size and Interpretation of the European Organisation for the Research and Treatment of Cancer Quality of Life Questionnaire Core 30" *Journal of Clinical Oncology (Jan 2011), Number 29, Issue 1, p89-96.*

The results in Section 7.1 are based on this publication but have been re-written to include more information on the results from the meta-analysis than could be included in the published article.

7.1 Cross-sectional contrasts

7.1.1 Number of contrasts

There were 2212 cross-sectional contrasts in the analysis dataset. Within the different subscales this led to a sample size of between 0 and 102 contrast available for the estimates of trivial, small, medium and large effects (Table 52). Although the initial number of contrasts in the dataset was very large, there were only 30 contrasts across all of the subscales which were classed as large differences by the expert review process. The number of contrasts was therefore generally too low to reliably estimate large effects by subscale (zero to six contrasts) so meta-analyses were not carried out for the large size class. There were also fewer than five contrasts for the medium size class in the DI and FI subscales therefore estimates for these were also excluded.

Table 52 Number of contrasts for estimates of cross-sectional effects

Expert size class	Subscale															Total
	AP	CF	CO	DI	DY	EF	FA	FI	NV	PA	PF	QL	RF	SF	SL	
Trivial	62	70	72	96	62	54	38	54	58	53	32	39	40	42	47	819
Small	56	74	44	38	38	91	71	25	56	62	102	97	96	101	57	1008
Medium	14	19	11	1	12	33	39	3	17	30	48	54	32	36	6	355
Large	2	0	1	0	0	4	2	0	0	3	6	5	3	3	1	30
Total	134	163	128	135	112	182	150	82	131	148	188	195	171	182	111	2212

7.1.2 Meta-analysis of mean differences

7.1.2.1 Estimates of random effects

The random effects and 95% confidence intervals were estimated using the residual (restricted) maximum likelihood (REML) method (using SAS® PROC MIXED). These were then used in the weighting of the contrasts in the meta-analysis (see Section 4.4.2.3 for details). The mean estimate of the random effect was used for the main analysis, with sensitivity analyses carried out using the upper and lower 95% confidence intervals for the estimate instead. The size of the estimated random effects variance gives an indication of the degree of heterogeneity of the mean differences across contrasts. The lowest estimate was for the DI subscale (14.9) and the highest estimate of the random effect was for RF (80.0).

7.1.2.2 Random effects models results

Table 53 shows the results from the meta-analysis of mean differences from the cross-sectional contrasts. The table shows the weighted mean differences from the random effects model along with 95% confidence intervals for each expert size class within a subscale.

The results are also shown in a box and whisker plot (Figure 44). There are three plots for each subscale, representing trivial, small and medium estimates from left to right. An 'X' has been used to show where there is no data to display for a certain size class for that subscale. Each subscale is shown with a different plot symbol (see key below the graph).

Most subscales showed clear trends across size classes, with an increase in estimates from trivial through to medium. Role functioning showed the widest range between the estimates (0, 13 and 25 points for trivial, small and medium size classes respectively), while other subscales, such as global quality of life had a smaller range

of estimates between trivial and medium size classes (1, 7 and 13). For the EF, SF, CF, CO and DY subscales there was some degree of overlap between the confidence intervals for the small and medium size classes. Of particular concern was the EF subscale which had a higher estimate for small (8.3) than for the medium size class (6.5). For this reason the subscale was not taken forward to derive the guidelines. (See 7.4 for further discussion.) Estimates for average trivial effects were at most 1.1 points (PF). Estimates for average small effects ranged from 4.7 to 12.7 points (for NV and RF subscales respectively). Estimates for average medium effects ranged from 10.1 (DY) to 25.1 points (RF).

Note for the trivial size class there were some estimates above and some below zero, however all of the confidence intervals span zero as one would expect for this category. A negative mean difference estimate means that on average Group 2 was worse than Group 1. However, as discussed in Section 4.4.1.2.2, the allocation of Group 1 and Group 2 had arbitrary meaning for the majority of contrasts therefore a slightly negative or positive mean difference simply means the absolute magnitude of the mean differences was slightly different to zero on average.

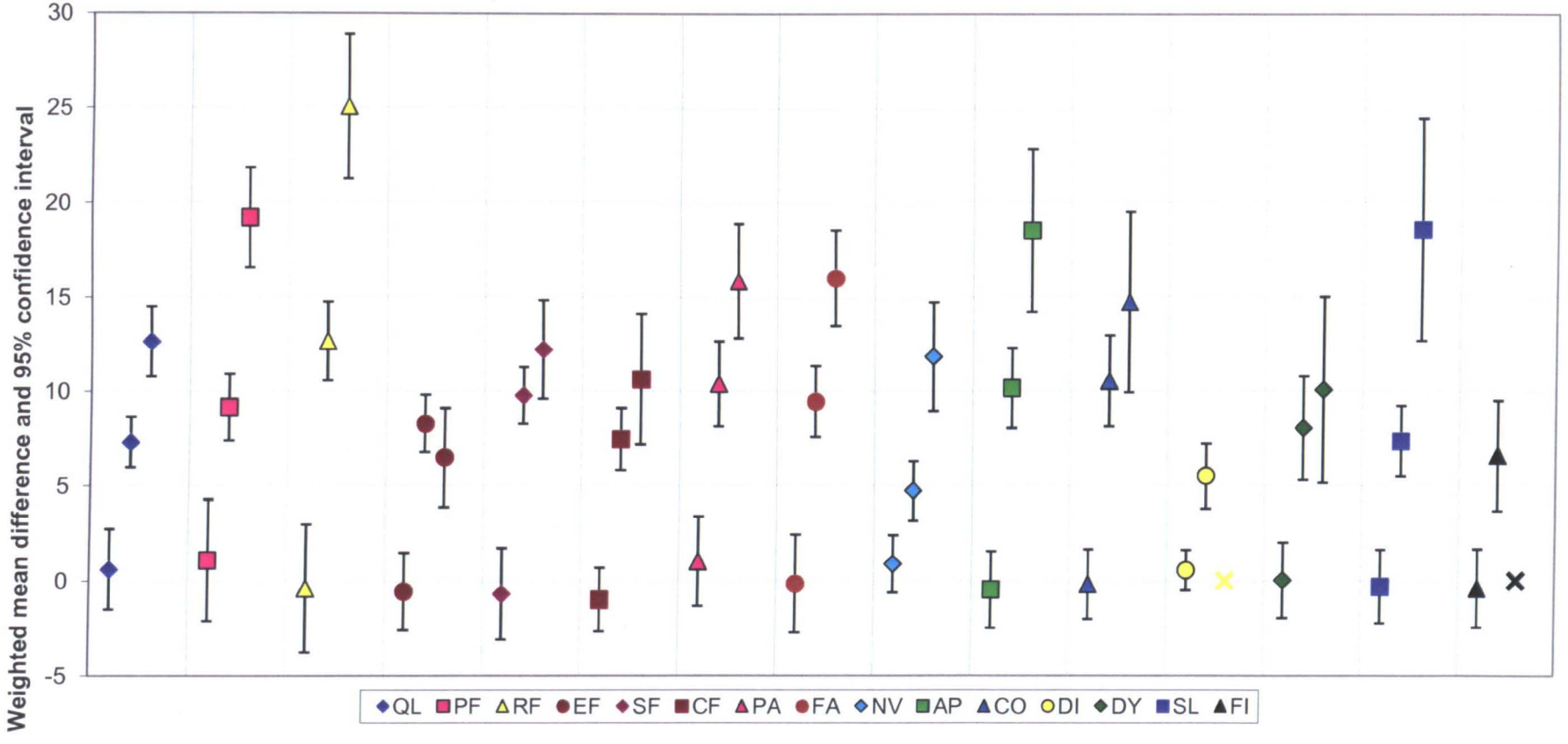
Table 53 Estimates for mean difference outcome variable by size category (cross-sectional contrasts)

Subscale	Expert size class	Weighted effect size	Lower confidence interval	Upper confidence interval	Number of contrasts
CF	Trivial	-0.98	-2.65	0.70	39
CF	Small	7.47	5.82	9.12	97
CF	Medium	10.64	7.19	14.09	54
EF	Trivial	-0.54	-2.58	1.49	32
EF	Small	8.30	6.77	9.82	102
EF	Medium	6.50	3.89	9.12	48
PA	Trivial	1.04	-1.31	3.40	40
PA	Small	10.40	8.14	12.66	96
PA	Medium	15.86	12.83	18.89	32
PF	Trivial	1.11	-2.09	4.30	54
PF	Small	9.18	7.41	10.94	91
PF	Medium	19.21	16.58	21.85	33
QL	Trivial	0.63	-1.49	2.75	42
QL	Small	7.33	5.99	8.67	101

Subscale	Expert size class	Weighted effect size	Lower confidence interval	Upper confidence interval	Number of contrasts
QL	Medium	12.67	10.82	14.52	36
RF	Trivial	-0.37	-3.74	2.99	70
RF	Small	12.67	10.58	14.76	74
RF	Medium	25.08	21.26	28.89	19
SF	Trivial	-0.67	-3.08	1.74	53
SF	Small	9.78	8.27	11.29	62
SF	Medium	12.23	9.61	14.84	30
FA	Trivial	-0.14	-2.71	2.44	38
FA	Small	9.48	7.59	11.36	71
FA	Medium	16.02	13.48	18.57	39
NV	Trivial	0.90	-0.61	2.41	58
NV	Small	4.73	3.16	6.30	56
NV	Medium	11.85	8.97	14.73	17
AP	Trivial	-0.45	-2.47	1.57	62
AP	Small	10.18	8.05	12.30	56
AP	Medium	18.52	14.20	22.84	14
CO	Trivial	-0.16	-2.00	1.67	72
CO	Small	10.56	8.15	12.97	44
CO	Medium	14.73	9.96	19.50	11
DI	Trivial	0.57	-0.49	1.64	96
DI	Small	5.52	3.81	7.24	38
DI	*Medium*				*1**
DY	Trivial	0.05	-1.93	2.03	62
DY	Small	8.06	5.30	10.82	38
DY	Medium	10.08	5.15	15.01	12
SL	Trivial	-0.28	-2.21	1.65	47
SL	Small	7.38	5.51	9.25	57
SL	Medium	18.57	12.68	24.45	6
FI	Trivial	-0.38	-2.44	1.67	54
FI	Small	6.60	3.68	9.51	25
FI	*Medium*				*3**

*Size classes with number of contrasts less than 5 not included in the guidelines

Figure 44 Estimates for mean difference outcome variable by expert size class (cross-sectional contrasts)



Graphs from left to right for each subscale represent estimates for trivial, small and medium size classes respectively

7.1.3 Meta-analysis of effect sizes

7.1.3.1 Estimates of random effects

The random effects and 95% confidence intervals were estimated using the residual (restricted) maximum likelihood (REML) method (using SAS® PROC MIXED). The mean estimate of the random effect was used in the weighting of the contrasts in the meta-analysis (see Section 4.4.2.3 for details). Note that sensitivity analyses regarding the use of the mean random effects variance for the effect size outcome variable were not carried out separately from those for the mean difference outcome variable. The sensitivity analyses for the mean difference outcome variable should indicate the robustness of results compared to using the upper/lower 95% confidence limits of the random effects estimates.

There were problems in estimating the random effects for the effect size outcome variable, with models not converging in four of the subscales. This can be caused by small numbers within a group. However, even after excluding any size classes with very small numbers of contrasts the models still had convergence issues. One of the other issues that can cause convergence problems in SAS® PROC MIXED is working with small numbers (i.e. where the outcome variable is on a small scale rather than number of contrasts being small). Since effect sizes are relatively small numbers (i.e. generally less than one) I tried rescaling the effect sizes using a multiplier of 100 in order to run the models and then adjusted the results back to the original scale. Note to adjust the variances back to the original scale a divisor of (100×100) was thus required. This resolved any problems with convergence and makes no difference to the resulting random effect estimates.

The size of the estimated random effects variance gives an indication of the degree of heterogeneity of the effect sizes across contrasts. As would be expected, the observations made on the mean difference outcome variable also held true here, i.e. that DI had the lowest estimate of random effects variance (0.05) and RF had the highest (0.39).

7.1.3.2 Results from random effects models

Table 54 shows the results from the meta-analysis of effect sizes from the cross-sectional contrasts. The table shows the weighted effect sizes from the random effects

models along with 95% confidence intervals for each expert size class within a subscale.

The results are also shown in a box and whisker plot (Figure 45). There are three plots for each subscale, representing trivial, small and medium estimates from left to right. Each subscale is shown with a different plot symbol (see key below the graph).

Within most subscales the increase in estimates was clear from trivial through to medium size classes. Role functioning had the widest range of estimates from zero for trivial, 0.5 for small and 0.8 for medium. The QL subscale in comparison had a much smaller range (although the confidence intervals do not overlap between the size classes) with zero for trivial, 0.3 for small and 0.5 for medium size classes. Confidence intervals for the small and medium estimates were overlapping for EF, SF, CF, CO and DY subscales. EF also had a higher estimate for the small size class than for the medium size class so was excluded from the guidelines.

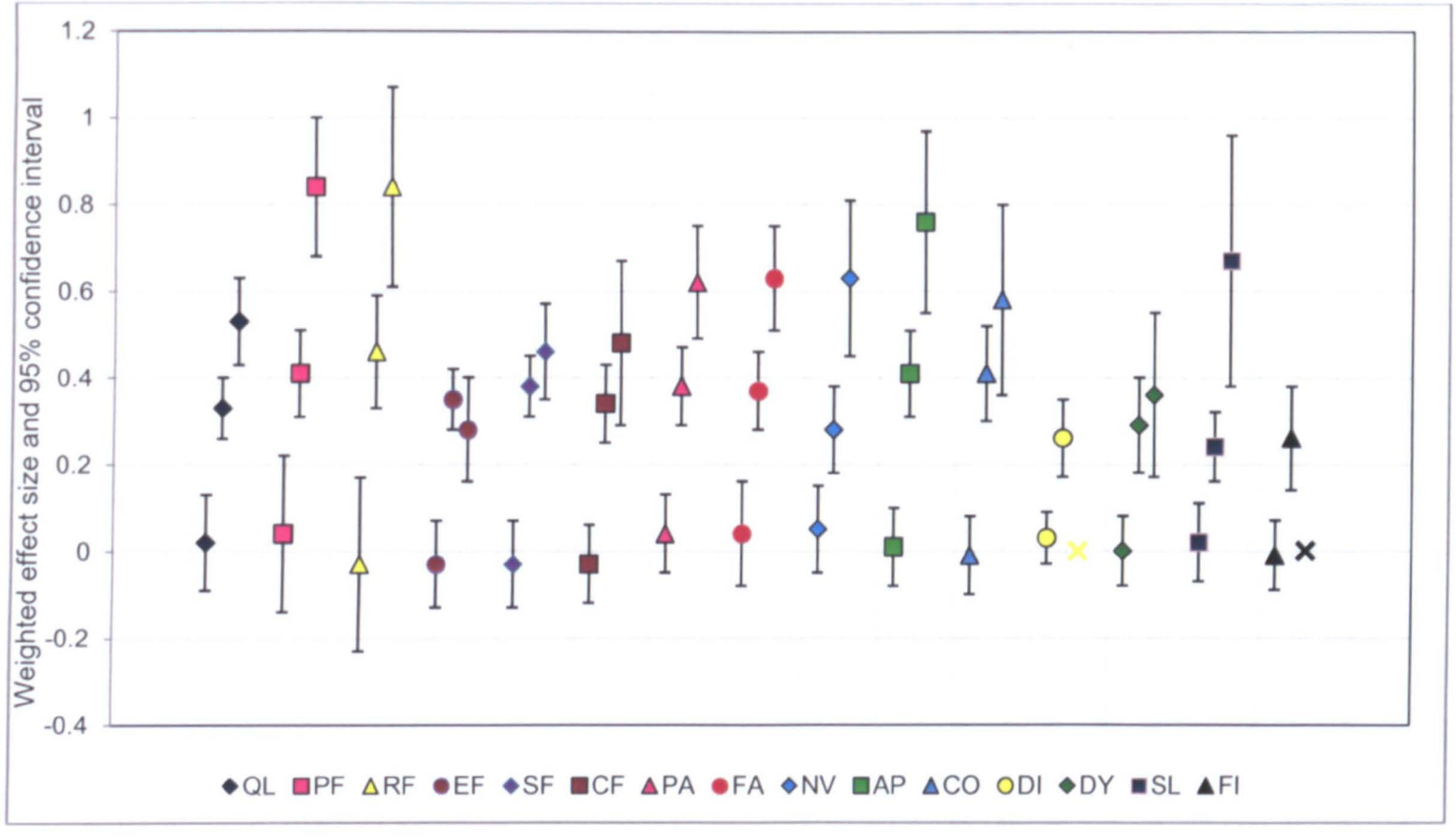
Table 54 Estimates for effect size outcome variable by expert size class (cross-sectional contrasts)

Subscale	Expert size class	Weighted effect size	Lower confidence interval	Upper confidence interval	Number of contrasts
QL	Trivial	0.02	-0.09	0.13	39
QL	Small	0.33	0.26	0.40	97
QL	Medium	0.53	0.43	0.62	54
PF	Trivial	0.04	-0.14	0.22	32
PF	Small	0.41	0.31	0.51	102
PF	Medium	0.84	0.68	0.99	48
RF	Trivial	-0.03	-0.23	0.18	40
RF	Small	0.46	0.33	0.60	96
RF	Medium	0.84	0.61	1.07	32
EF	Trivial	-0.03	-0.13	0.07	54
EF	Small	0.35	0.28	0.43	91
EF	Medium	0.28	0.16	0.41	33
SF	Trivial	-0.03	-0.13	0.07	42
SF	Small	0.38	0.31	0.44	101
SF	Medium	0.46	0.35	0.57	36
CF	Trivial	-0.03	-0.12	0.06	70

Subscale	Expert size class	Weighted effect size	Lower confidence interval	Upper confidence interval	Number of contrasts
CF	Small	0.34	0.25	0.43	74
CF	Medium	0.48	0.29	0.66	19
PA	Trivial	0.04	-0.05	0.14	53
PA	Small	0.38	0.29	0.48	62
PA	Medium	0.62	0.49	0.75	30
FA	Trivial	0.04	-0.08	0.17	38
FA	Small	0.37	0.28	0.47	71
FA	Medium	0.63	0.51	0.76	39
NV	Trivial	0.05	-0.05	0.15	58
NV	Small	0.28	0.18	0.38	56
NV	Medium	0.63	0.45	0.81	17
AP	Trivial	0.01	-0.08	0.11	62
AP	Small	0.41	0.31	0.51	56
AP	Medium	0.76	0.55	0.96	14
CO	Trivial	-0.01	-0.1	0.07	72
CO	Small	0.41	0.3	0.52	44
CO	Medium	0.58	0.36	0.79	11
DI	Trivial	0.03	-0.03	0.09	96
DI	Small	0.26	0.17	0.35	38
DI	*Medium*				*1**
DY	Trivial	0	-0.08	0.08	62
DY	Small	0.29	0.18	0.40	38
DY	Medium	0.36	0.17	0.55	12
SL	Trivial	0.02	-0.07	0.11	47
SL	Small	0.24	0.16	0.33	57
SL	Medium	0.67	0.38	0.95	6
FI	Trivial	-0.01	-0.09	0.07	54
FI	Small	0.26	0.14	0.38	25
FI	*Medium*				*3**

*Size classes with number of contrasts less than 5 not included in the guidelines

Figure 45 Estimates for effect size outcome variable by expert size class (cross-sectional contrasts)



Graphs from left to right for each subscale represent trivial, small and medium estimates respectively

7.1.4 Guidelines for comparing between groups of patients

The resulting guidelines for trivial, small, medium and large effects are provided in Table 55 for both mean differences and effect sizes. Section 4.5 describes how the guidelines were calculated from the meta-analysis weighted estimates. Although there was insufficient data to estimate the size of large effects, the upper limit of the 95% confidence intervals around the medium estimates have been used as a guide.

The method of obtaining the guidelines is illustrated here using the global quality of life scale as an example. Mean estimates for trivial, small and medium mean size classes were 1, 7 and 13 points respectively (Table 53). The threshold between size classes was set at the midpoint between estimates. Therefore the threshold between trivial and small is 4 points, i.e. the midpoint between 1 and 7. The threshold between small and medium differences is 10 points, i.e. at the midpoint between 7 and 13. The threshold between medium and large differences cannot be determined as there was insufficient data to obtain an estimate for the large size class. Instead the threshold is set at the upper limit of the 95% confidence interval around the medium size class estimate, i.e. at 15 points.

Table 55 Guidelines for size of cross-sectional differences (from meta-analysis)

Threshold between small and medium estimates	Sub-scale	Mean difference				Effect size			
		Triv	Small	Medium	Large	Triv	Small	Medium	Large
<10 points	DI	0 - 3	3- 7	>7	-	0-0.1	0.1-0.4	>0.4	-
	NV	0 - 3	3- 8	8 - 15	>15	0-0.2	0.2-0.5	0.5-0.8	>0.8
	CF	0 - 3	3 - 9	9 - 14	>14	0-0.2	0.2-0.4	0.4-0.7	>0.7
	DY	0 - 4	4 - 9	9 - 15	>15	0-0.1	0.1-0.3	0.3-0.6	>0.6
10-15 points	FI	0 - 3	3 - 10	>10	-	0-0.1	0.1-0.4	>0.4	-
	QL	0 - 4	4 - 10	10 - 15	>15	0-0.2	0.2-0.4	0.4-0.6	>0.6
	SF	0 - 5	5 - 11	11 - 15	>15	0-0.2	0.2-0.4	0.4-0.6	>0.6
	SL	0 - 4	4 - 13	13 - 24	>24	0-0.1	0.1-0.5	0.5-1	>1
	FA	0 - 5	5 - 13	13 - 19	>19	0-0.2	0.2-0.5	0.5-0.8	>0.8
	CO	0 - 5	5 - 13	13 - 19	>19	0-0.2	0.2-0.5	0.5-0.8	>0.8
	PA	0 - 6	6 - 13	13 - 19	>19	0-0.2	0.2-0.5	0.5-0.8	>0.8
	PF	0 - 5	5 - 14	14 - 22	>22	0-0.2	0.2-0.6	0.6-1	>1
	AP	0 - 5	5 - 14	14 - 23	>23	0-0.2	0.2-0.6	0.6-1	>1
>15 points	RF	0 - 6	6 - 19	19 - 29	>29	0-0.2	0.2-0.7	0.7-1.1	>1.1

In order to use these guidelines to calculate a sample size (assuming it is required to detect the smallest clinically relevant difference) the threshold between trivial and small should be used in the calculation, i.e. 4 points for the QL subscale, 5 points for the PF subscale and so on. To use the guidelines for interpretation, an observed difference of 5 points for example would be interpreted as trivial for the PA and RF subscales but would lie in the small range for all other subscales. Note that where the threshold between trivial and small is 5 points this is interpreted as a small difference.

7.2 Longitudinal contrasts

7.2.1 Number of contrasts

There were 1232 contrasts in the analysis dataset which compared groups over time. For these contrasts there were more groupings from the expert size classes than for the cross-sectional contrasts, since the direction of the difference (improving or declining over time) is meaningful. Due to the increased number of size classes and also the agreement criteria reducing the available contrasts, the numbers of longitudinal contrasts was quite small when split into subscales and then size classes within the subscales (Table 56).

As with the cross-sectional contrasts, the number of contrasts judged as a large difference in either direction was very low, even before applying the agreement criteria. There were only three contrasts deemed large in the analysis dataset. The number of contrasts deemed medium in either direction was also low (ranging from zero to 24 across subscales). The number of contrasts meeting the agreement criteria was particularly low for the trivial size class (i.e. for some subscales there were very few contrasts where all experts were in agreement that the difference would be trivial). There were five subscales where an estimate from meta-analysis was not obtained for the trivial size class as the number of contrasts was less than five.

Table 56 Number of contrasts for estimates of longitudinal effects

Expert size class	Subscale															Total
	AP	CF	CO	DI	DY	EF	FA	FI	NV	PA	PF	QL	RF	SF	SL	
Large deterioration	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	2
Medium deterioration	7	1	3	4	2	2	24	2	16	5	15	16	13	10	5	125
Small deterioration	25	28	11	14	14	11	40	9	27	17	39	30	35	38	12	350
Trivial	21	17	37	46	21	6	4	22	14	11	2	1	2	3	10	217
Small improvement	20	20	17	10	21	87	28	4	20	50	27	59	35	41	35	474
Medium improvement	3	0	2	1	0	9	6	0	2	9	7	12	4	5	3	63

Expert size class	Subscale															
	AP	CF	CO	DI	DY	EF	FA	FI	NV	PA	PF	QL	RF	SF	SL	Total
Large improvement	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
Total	76	66	70	75	58	116	102	37	80	92	91	118	89	97	65	1232

7.2.2 Meta-analysis of mean differences

7.2.2.1 Estimates of random effects

The random effects and 95% confidence intervals were estimated using the residual (restricted) maximum likelihood (REML) method (using SAS® PROC MIXED). The mean estimate of the random effect was used in the weighting of the contrasts in the meta-analysis (see Section 4.4.2.3 for details) and sensitivity analyses were used to check for the robustness of the results to using the mean estimate.

There were several problems in estimating the random effects variance using the REML method in SAS® PROC MIXED. A number of the models did not converge. Initially I treated each subscale independently and removed any size classes that had a small number of contrasts in them. (For cross-sectional models I had excluded where the number of contrasts was less than five and here I also tried where there were less than ten contrasts if convergence problems remained.) The subscales had to be treated individually as it was no longer the case that only the large size classes had small numbers in them. For some subscales the trivial and medium size classes were also too small but this did not apply universally across subscales. There were still convergence issues with four subscales (CF, EF, PF and QL) indicating that there was probably another reason for the convergence problem. As for the effect size analysis within the cross-sectional results I tried re-scaling the outcome variable instead. The longitudinal mean differences were fairly small in comparison to the cross-sectional ones therefore it was possible that this was also causing problems with the model convergence. I multiplied the mean differences by a factor of 10 and then ran the models to estimate the random effects variances. All of the models then converged, indicating that the overriding factor here was the small scale of the analysis variable rather than small numbers of contrasts. The variances were re-scaled back to the original scale by dividing by 100, i.e. 10×10).

The AP subscale had the highest random effect estimate (77.3) and FI had the lowest (24.1). This differs from the cross-sectional analysis where RF had the highest and DI the lowest estimate.

7.2.2.2 Results from random effects models

Table 57 shows the results from the random effects models. The table shows the weighted mean differences along with 95% confidence intervals for each expert size class within a subscale.

The results are also shown in a box and whisker plot (Figure 46). There are five plots for each subscale, representing medium deterioration, small deterioration, trivial, small improvement and medium improvement from left to right. Each subscale is shown with a different plot symbol (see key below the graph). An 'X' has been used to show where there is no data to display for a certain size class for that subscale. Note that estimates where number of contrasts is less than 5 have been excluded in the graph.

Table 57 Estimates for mean difference outcome variable by expert size class (longitudinal contrasts)

Sub-scale	Expert size class	Weighted mean difference	Lower confidence interval	Upper confidence interval	Number of contrasts
QL	Medium deterioration	-12.89	-15.98	-9.80	16
QL	Small deterioration	-7.25	-9.47	-5.02	30
QL	Trivial				1*
QL	Small improvement	6.33	4.74	7.92	58
QL	Medium improvement	5.37	1.93	8.81	12
PF	Medium deterioration	-12.94	-16.50	-9.38	15
PF	Small deterioration	-7.73	-10.00	-5.46	37
PF	Trivial				2*
PF	Small improvement	4.86	2.24	7.47	27
PF	Medium improvement	4.33	-0.71	9.37	7

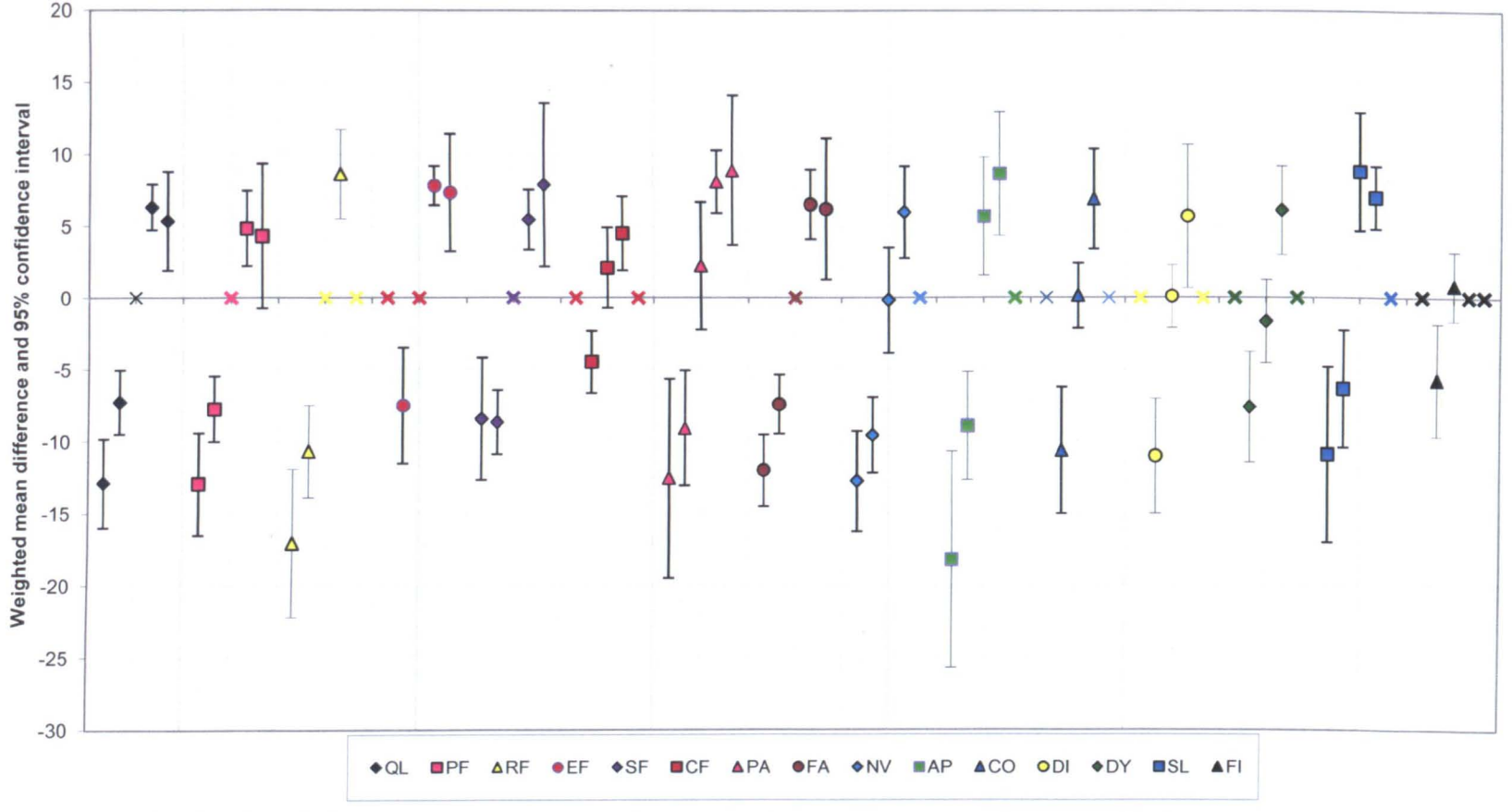
Sub-scale	Expert size class	Weighted mean difference	Lower confidence interval	Upper confidence interval	Number of contrasts
RF	Medium deterioration	-17.04	-22.18	-11.89	13
RF	Small deterioration	-10.68	-13.89	-7.48	33
RF	Trivial				2*
RF	Small improvement	8.65	5.55	11.76	35
RF	Medium improvement				4*
EF	Medium deterioration				2*
EF	Small deterioration	-7.50	-11.53	-3.47	11
EF	Trivial				4*
EF	Small improvement	7.85	6.49	9.2	86
EF	Medium improvement	7.37	3.28	11.46	9
SF	Medium deterioration	-8.44	-12.71	-4.18	10
SF	Small deterioration	-8.67	-10.91	-6.44	38
SF	Trivial				3*
SF	Small improvement	5.48	3.39	7.57	40
SF	Medium improvement	7.90	2.20	13.6	5
CF	Medium deterioration				1*
CF	Small deterioration	-4.47	-6.65	-2.29	28
CF	Trivial	2.12	-0.69	4.92	17
CF	Small improvement	4.52	1.93	7.10	20
CF	Medium improvement				0*
PA	Medium deterioration	-12.58	-19.5	-5.65	5
PA	Small deterioration	-9.08	-13.09	-5.06	17
PA	Trivial	2.26	-2.20	6.72	11

Sub-scale	Expert size class	Weighted mean difference	Lower confidence interval	Upper confidence interval	Number of contrasts
PA	Small improvement	8.13	5.95	10.32	50
PA	Medium improvement	8.93	3.72	14.15	9
FA	Medium deterioration	-12.02	-14.52	-9.51	24
FA	Small deterioration	-7.42	-9.47	-5.36	38
FA	Trivial				4*
FA	Small improvement	6.55	4.11	8.99	28
FA	Medium improvement	6.23	1.29	11.17	6
NV	Medium deterioration	-12.78	-16.28	-9.28	16
NV	Small deterioration	-9.56	-12.21	-6.91	27
NV	Trivial	-0.16	-3.85	3.52	14
NV	Small improvement	6.00	2.80	9.19	20
NV	Medium improvement				2*
AP	Medium deterioration	-18.21	-25.72	-10.71	7
AP	Small deterioration	-8.93	-12.7	-5.16	25
AP	Trivial	5.70	1.59	9.80	21
AP	Small improvement	8.68	4.36	13.00	20
AP	Medium improvement				3*
CO	Medium deterioration				3*
CO	Small deterioration	-10.64	-15.05	-6.23	11
CO	Trivial	0.16	-2.10	2.42	37
CO	Small improvement	6.91	3.43	10.40	17
CO	Medium improvement				2*

Sub-scale	Expert size class	Weighted mean difference	Lower confidence interval	Upper confidence interval	Number of contrasts
DI	Medium deterioration				4*
DI	Small deterioration	-11.03	-15.05	-7.02	14
DI	Trivial	0.11	-2.07	2.30	46
DI	Small improvement	5.70	0.70	10.70	10
DI	Medium improvement				1*
DY	Medium deterioration				2*
DY	Small deterioration	-7.60	-11.45	-3.75	14
DY	Trivial	-1.62	-4.51	1.27	21
DY	Small improvement	6.13	3.04	9.22	21
DY	Medium improvement				0*
SL	Medium deterioration	-10.89	-16.98	-4.81	5
SL	Small deterioration	-6.31	-10.41	-2.20	12
SL	Trivial	8.81	4.69	12.92	10
SL	Small improvement	7.00	4.81	9.19	35
SL	Medium improvement				3*
FI	Medium deterioration				2*
FI	Small deterioration	-5.71	-9.64	-1.77	9
FI	Trivial	0.83	-1.58	3.23	22
FI	Small improvement				4*
FI	Medium improvement				0*

*Size classes with number of contrasts less than 5 not included in the guidelines

Figure 46 Estimates for mean difference outcome variable by expert size class (longitudinal contrasts)



Graphs from left to right for each subscale represent medium deterioration, small deterioration, trivial, small improvement and medium improvement size classes respectively

There were less estimates obtainable through meta-analysis than for the cross-sectional contrasts. Only the PA subscale had an estimate for all of the size classes from medium deterioration up to a medium improvement. FI had the least available estimates with only two of the size classes with obtainable estimates.

Medium deteriorations in scores could be estimated in nine subscales and these all lay in the 10-20 points range except for the SF subscale where a medium deterioration was estimated at 8 points (although this estimate was only from 10 contrasts). Medium improvements were only obtainable in six subscales and ranged from 4 to 9 points. All subscales with estimates for both medium deteriorations and improvements available had deterioration estimates of a larger magnitude than the improvements. For example, for the PA subscale the medium deterioration estimate was -12.6 compared with the medium improvement estimate of 8.9.

Small deteriorations in scores could be estimated in all subscales. The estimates lay between 4 and 11 points. Small improvements could also be estimated in 14 of the subscales. The estimates for improvements were again lower in magnitude than their corresponding estimate for a small deterioration in score for 11 of the subscales. However, across most subscales there was some overlap of the confidence intervals between the small and medium estimates.

Estimates for trivial differences were possible for nine of the subscales. The estimates were close to zero with confidence intervals spanning zero for all subscales except the AP and SL subscales where trivial differences were estimated at 6 and 9 points respectively.

7.2.3 Guidelines for comparing groups of patients over time

As several of the size classes could not be estimated within subscales the derivation of guidelines was more difficult than for the cross-sectional contrasts. The same methodology was followed with some minor adjustments. As for the cross-sectional contrasts, the midpoint between two estimates was used as the threshold between size classes and 95% confidence intervals from the medium estimates were used to inform the guidelines for large size classes. However, in a number of subscales medium estimates were not available, in which case an estimate for the large class was unobtainable and the 95% confidence intervals around the small estimates were used to inform the size of medium size classes. There were a couple of subscales (RF and FI) where estimates were available for some size classes but not necessarily from

neighbouring size classes (e.g. for RF where estimates of small deteriorations and small improvements were possible but not for the trivial size class). Using midpoints between estimates meant that the guidelines could only be written for large and medium deteriorations, losing all of the information we had around the estimates for small size classes. For this reason I used the confidence intervals around the small estimates to add to the guidelines (e.g. the upper 95% confidence interval (CI) of the small deterioration and lower 95% CI of the small improvement class defined the trivial guidelines). In this way all of the information could be used and the guidelines could be more informative for the RF and FI subscales.

Table 58 shows the subscales with guidelines that can be derived from the meta-analysis estimates. (NE indicates a guideline for that size class was unobtainable.)

There were six subscales (QL, PF, EF, SF, FA and SL) where the ordering of estimates across size classes did not follow the expected gradient. For example, the QL subscale had a small estimate of 6.3 and a medium estimate of 5.4. Typically, the size classes that displayed this illogical ordering contained relatively few contrasts (six to 12). Although the team decided during the development of the methods to exclude estimates where the number of contrasts was less than 5, this was a fairly arbitrary decision and it may be that the common illogical ordering of longitudinal estimates indicates that more than 5 contrasts are required for a reliable estimate. For the purposes of developing guidelines for longitudinal results, I have used the information from the size class with the larger number of contrasts for these subscales. So for the above example in the QL subscale, I have used the estimate for the small size class (from 58 contrasts) and its 95% confidence interval to inform the guidelines rather than the medium estimate (which was only from 12 contrasts). This allowed guidelines to be provided for these important subscales.

Table 58 Guidelines for size of longitudinal differences (from meta-analysis)

Threshold between small and medium improvements (deteriorations)	Sub-scale	Deteriorations			No difference	Improvements		
		Large	Medium	Small	Trivial	Small	Medium	Large
<10 (>-14)	FI	NE	<-10	-10 to -2	-2 to 3	>3	NE	NE
	CF	NE	<-7	-7 to -1	-1 to 3	3 to 7	>7	NE
	PF	<-17	-17 to -10	-10 to -5	-5 to 2	2 to 7	>7	NE
	QL	<-16	-16 to -10	-10 to -5	-5 to 5	5 to 8	>8	NE
	SF	NE	<-11	-11 to -6	-6 to 3	3 to 8	>8	NE
	EF	NE	<-12	-12 to -3	-3 to 6	6 to 9	>9	NE
	NV	<-16	-16 to -11	-11 to -5	-5 to 3	3 to 9	>9	NE
	DY	NE	<-11	-11 to -5	-5 to 2	2 to 9	>9	NE
	FA	<-15	-15 to -10	-10 to -5	-5 to 4	4 to 9	>9	NE
	SL	<-17	-17 to -9	-9 to -2	-2 to 5	5 to 9	>9	NE
	PA	<-20	-20 to -11	-11 to -3	-3 to 5	5 to 9	9 to 14	>14
≥10 (≤-14)	CO	NE	<-15	-15 to -5	-5 to 4	4 to 10	>10	NE
	DI	NE	<-15	-15 to -5	-5 to 3	3 to 11	>11	NE
	RF	<-22	-22 to -14	-14 to -7	-7 to 6	6 to 12	>12	NE
	AP	<-26	-26 to -14	-14 to -2	-2 to 7	7 to 13	>13	NE

To use these guidelines for interpretation, an observed improvement of five points would be classed as trivial for AP, RF and EF subscales, whereas for the QL, SF, CO, CF, DI, DY, FA, NV and PA subscales a difference of this size would be classed as small. If there was a deterioration over time of the same magnitude (i.e. 5 points) then it would only be considered trivial for the RF and SF scales; for the remaining subscales it would represent a small deterioration.

In order to use these guidelines for a sample size calculation (assuming it is required to detect the smallest clinically relevant improvement) the threshold between trivial and small should be used in the calculation.

7.3 Results from sensitivity analyses

7.3.1 Imputed variance estimates

A hierarchical approach to obtaining the required variances for the meta-analysis was used. Where possible variances were derived using other information in the paper such as p-values (see details in Section 4.4.1.1), followed by methods requiring an assumed correlation value or, finally, single imputation by subscale using the reported variance estimates from other studies. A sensitivity analysis was carried out grouping the contrasts by the method of obtaining variances. This was carried out across subscales and expert size class in order to give an overall indication of the weighted mean estimates for each method.

For the cross-sectional contrasts, 57% of contrasts had variance data available directly or derived using other data in the original article (Table 59). The remaining contrasts (43%) had variance data imputed using single imputation. This is probably analogous to standard meta-analyses where less than half of the studies may report the necessary variances for analysis (72). There were substantially less longitudinal contrasts with variance data available from the original article (8%). Around 40% of contrasts had variances imputed using an assumed correlation coefficient of 0.5 (i.e. a mean change and the standard deviation for the change were reported but not standard deviations at the individual times). 50% needed single imputation. This is probably a much higher level of missing variance data than found in standard meta-analyses. However, here we were looking at comparisons over time whereas standard meta-analyses tend to look at between group comparisons. We also took longitudinal contrasts from articles where data over time was being reported (e.g. in a summary table or graph) but these were not necessarily the comparisons being focussed on in the article therefore a lower proportion of variances available was to be expected.

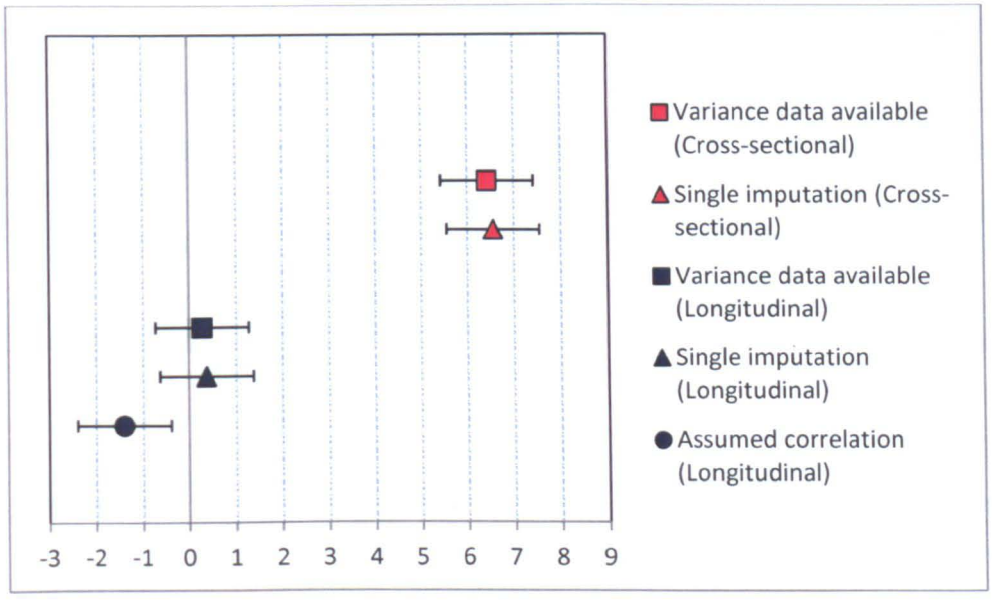
Table 59 Number of contrasts with available or imputed variance data

Number of contrasts (%)	Cross-sectional	Longitudinal
Variance data available	1252 (56.6%)	101 (8.2%)
Variance data through p	21 (0.9%)	5 (0.4%)
Assumed correlation value	0 (0.0%)	507 (41.2%)
Single imputation	939 (42.5%)	619 (50.2%)
Total	2212	1232

Figure 47 shows the meta-analysis estimates with 95% confidence intervals for contrasts from each of these subsets. Note, the number of contrasts imputed using p-values were too small to obtain separate estimates and are not displayed here. Estimates for contrasts with variance data available and for those where single imputation was used were very similar for both the longitudinal and cross-sectional contrasts. However, the longitudinal contrasts where an assumed correlation coefficient was used had a lower estimate. On average these were deteriorations in QOL whereas those with data available were improvements in QOL.

An estimate of the correlation coefficient (rho) is required for the analysis where a contrast has the mean change and standard deviation for the change reported rather than the standard deviations at each time point. Regardless of why the contrasts with data available or not are different with respect to the mean difference, it shows that it is important to include contrasts without full data available (otherwise the results would be biased) and therefore an estimate of the correlation coefficient is necessary. A sensitivity analysis was therefore carried out around the choice of correlation coefficient used in the analysis.

Figure 47 Forest plot for weighted mean difference grouped by method of obtaining variance for analysis



The meta-analysis was carried out assuming a correlation coefficient of 0.5 for the relationship between measurements on the same patient(70). I carried out sensitivity analyses by creating extra datasets where the imputation used correlation coefficients of 0.25 and 0.75 instead. The meta-analyses were repeated using these datasets and the results compared to see how much they varied. The median difference was 0 (range 0 to 0.13). The majority of estimates varied by 0 to 0.02 points.

The only differences larger then this were medium improvement for PA subscale and small deterioration for the FI subscale with differences of 0.12 and 0.13 respectively. These both had a relatively small number of contrasts (n=9) and if these consist mostly of contrasts requiring the assumption of a value for rho, this would explain why the change in assumption has a larger impact. However, changes even at this level would make no difference to the resulting guidelines.

7.3.2 Comparison of results from full dataset

The criteria for defining the analysis dataset were defined post hoc therefore the results from the full dataset are also displayed here for clarity.

These results confirm that the full dataset cannot be used to derive valid and useful estimates for the guidelines. There needs to be some rules applied so that the estimates from the meta-analysis represent the modulus of the combined contrasts rather than a diluted average arising from including estimates with uncertainty around the expert review or results.

Table 60 shows the proportion of contrasts from the full dataset that were used in the analysis dataset. The proportion used was higher for the medium size classes than for the small size classes. The proportion was lowest for the trivial difference but this is not surprising as the agreement criteria applied to the trivial size class was different to the other size classes and was quite strict.

Table 60 Proportion of cross-sectional contrasts included in analysis dataset by expert size class

Contrast type	Proportion of contrasts in full dataset included in analysis dataset (%)	Total
Cross-sectional	Trivial	36%
	Small	59%
	Medium	79%
Longitudinal	Medium deterioration	84%
	Small deterioration	64%
	Trivial	12%
	Small improvement	56%
	Medium improvement	80%

Figure 48 shows the results from the meta-analysis using all cross-sectional contrasts from the full dataset (i.e. those contrasts with at least two expert reviews).

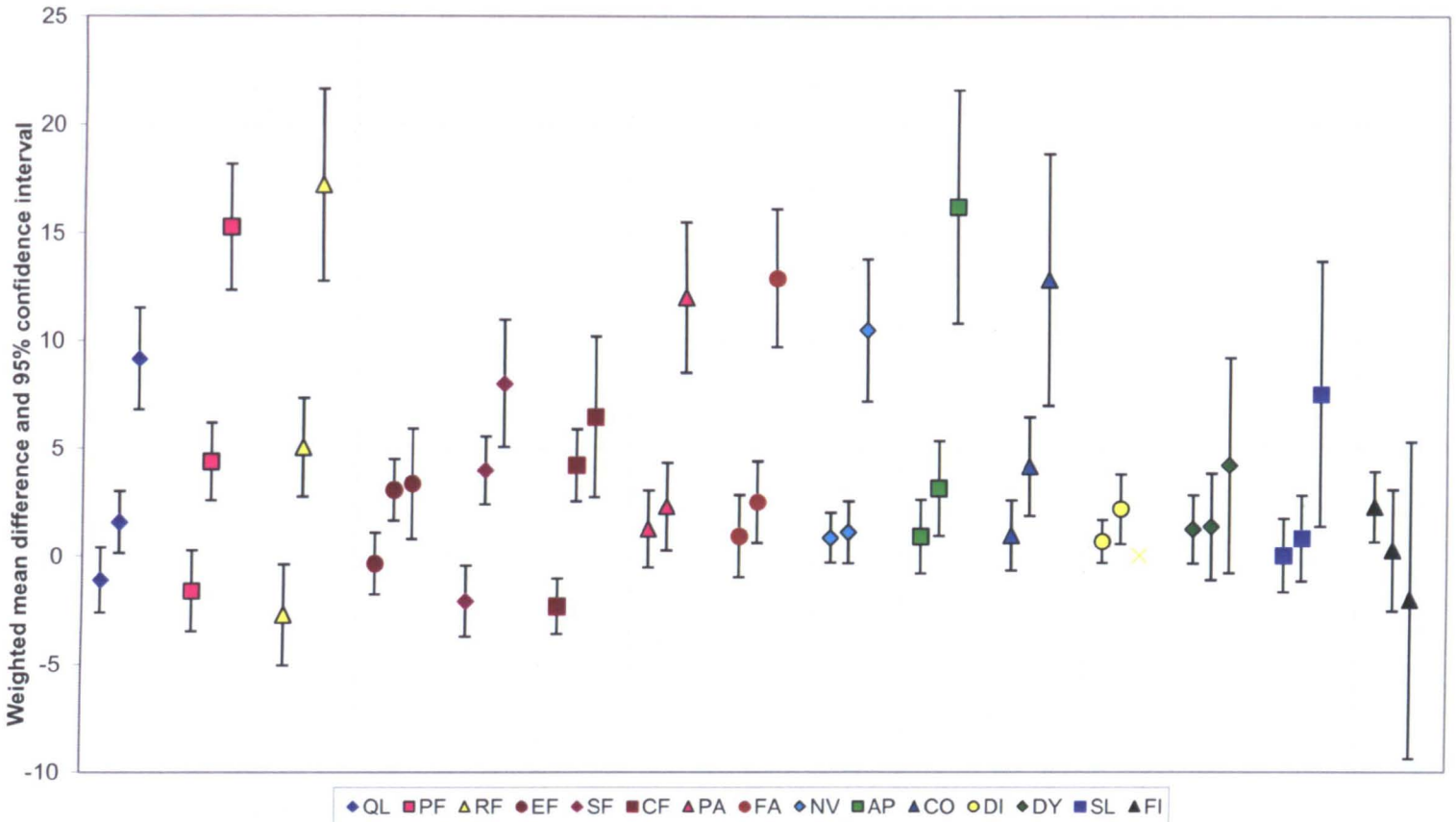
Compared to the main results from the analysis subset, the estimates for medium differences were lower (around 5 points lower on average) using the full dataset. The estimates for small differences were also smaller, they were all between zero and five in the full dataset, compared to 5 to 13 in the analysis dataset. The trivial estimates and confidence intervals no longer all span zero (three subscales were below and one subscale above zero in using the full dataset).

The small and medium estimates were lower from the full dataset because contrasts were included where the experts thought the difference would be in favour of one group whereas the observed mean difference showed a difference in the opposite direction (and hence would be a negative mean difference). By including all contrasts in the meta-analysis to get an estimate for small and medium contrasts these negative mean differences combined with the rest of the positive scores had the net result seen of reducing the estimates.

In terms of the trivial estimates, the full dataset contained contrasts where the experts did not agree that the contrast was a trivial one. These contrasts would be placed in the trivial size class because the average expert review was between -0.5 and 0.5 rather than because the consensus was that the effect was trivial. This confuses the estimates of trivial as seen here as some with reasonable sized mean differences have been included which skew the estimate.

These factors combine to make the trivial and small estimates largely indistinguishable from each other, i.e. the confidence intervals overlap. As an extreme example, there is a contrast in the full dataset where one expert had an average review of -3 and one had a review of +2.3. Therefore both reviewers considered the mean difference would be of a reasonable size (medium or large) but either they disagreed on the direction of the difference or one reviewer made an error when assigning the direction to their score. This contrast would be placed in the trivial size class as the average of -3 and 2.3 is -0.35 (i.e. between -0.5 and 0.5). However, when we look at the mean difference between the groups it is actually very high (42 points) and clearly does not belong in the trivial size class.

Figure 48 Meta-analysis results using full dataset – cross-sectional contrasts



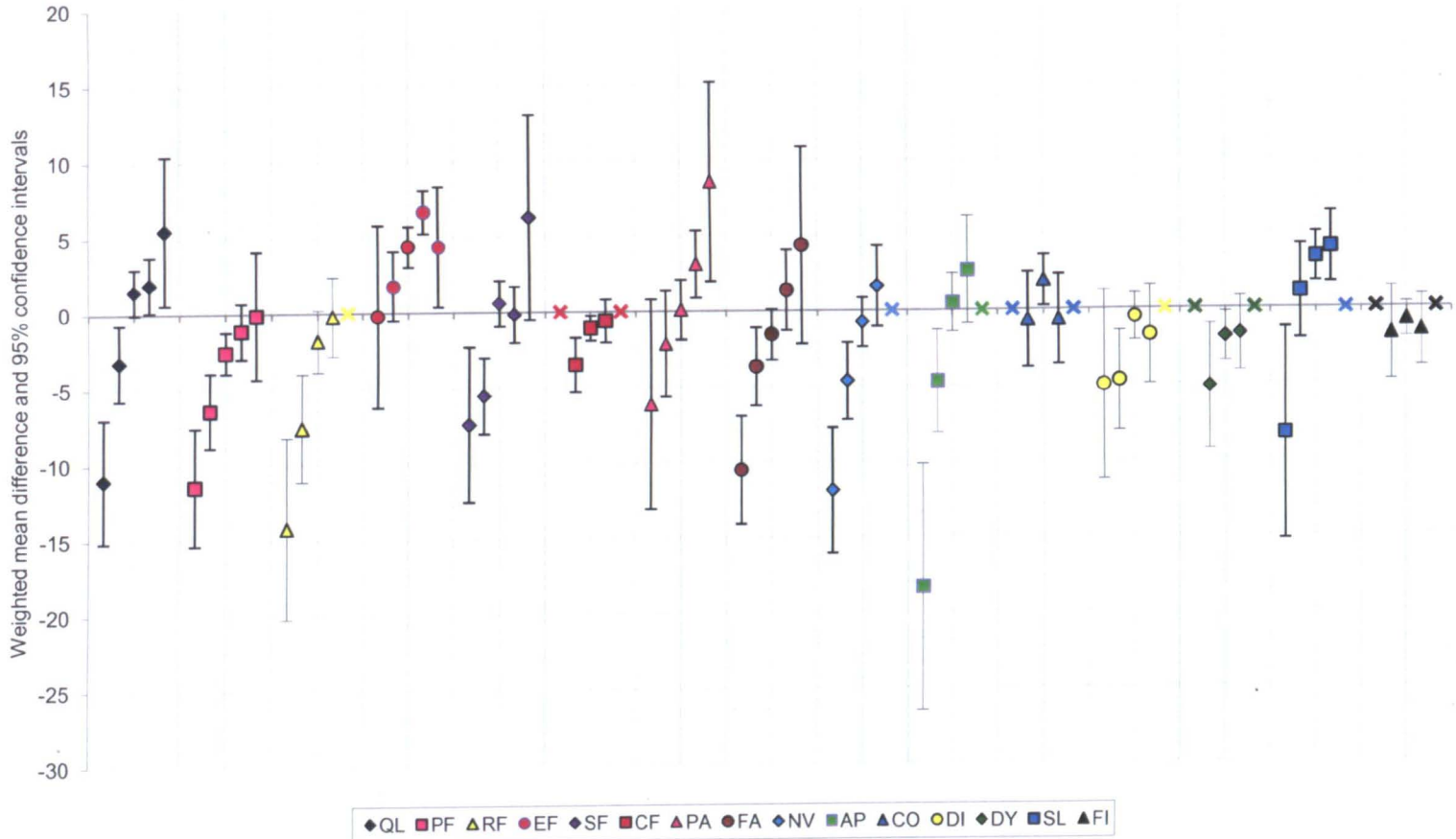
The three graphs for each subscale represent Trivial, Small and Moderate estimates from left to right

Figure 49 shows the results from the random effects meta-analysis including all longitudinal contrasts from the full dataset (i.e. contrasts with more than one expert review available). For five subscales the confidence intervals around the trivial estimates do not span zero. Table 60 shows that only 12% of the contrasts in the trivial size class could be used in the analysis dataset therefore estimates from the full dataset are likely to be very unreliable as a large proportion of contrasts were placed in the trivial size class due to averaging the experts judgements rather than all experts agreeing the contrast belonged in the trivial category It is likely that a large proportion of the contrasts in the full dataset that were classed as trivial actually contained reasonably sized mean differences.

We can also see that for 10 subscales the estimates and confidence intervals for small and even medium improvements are below or around zero, whereas for the analysis dataset the majority of estimates in these classes were above zero. Using PF as an example of one of these subscales we can see that none of the estimates are above zero. This implies that there are a number of contrasts where the experts judged a difference would be an improvement when actually the observed scores from the paper showed a deterioration or no difference. Including these contrasts diluted the estimates where there was good agreement between the experts and the original article, e.g. a mean difference of 10 points judged to be a medium improvement by the experts would be pooled with a mean difference of -10 points also judged to be a medium improvement by the experts. This had the net effect of zero points as an estimate for medium improvements when actually the modulus of the effect is around 10 points.

The agreement criteria used to reduce the full dataset to the analysis dataset were chosen such that contrasts where the expert size class and observed scores disagreed in direction were removed from the meta-analysis to avoid this dilution of the estimates. As Table 60 shows, just over half of the contrasts in the small improvement size class and 80% of the medium improvement contrasts met the criteria for the analysis dataset. A comparison of results from the full versus the analysis datasets, for these categories in particular, indicated that contrasts where the experts disagreed with the direction of the observed effect had a big difference on the resulting meta-analysis estimates, some of which did not make sense (such as a medium improvement estimate below zero).

Figure 49 Meta-analysis results using full dataset – longitudinal contrasts



7.3.3 Impact of using different estimates of the random effects variance

The main analysis results use an estimate of the average random effects variance for the random element of the weighting of contrasts in the meta-analysis. Sensitivity analyses were carried out to see how much the results varied when different estimates for the random effects variance were used. The upper and lower 95% confidence limits for the random effects variance were used to check how much the results could change if our estimate of the random effects variance was at either extreme rather than at the average.

For the cross-sectional analysis, changing the random element of the weighting affected the meta-analysis results only slightly, i.e. by an average of 0.05 points. In the worst cases the estimate changed by 0.2 points (PA subscale in the medium size class and DY subscale in the trivial size class). The midpoints between estimates were calculated for each scenario and there were no differences compared with the midpoints used to create the guidelines reported in Section 7.1.4.

For the longitudinal analysis the sensitivity analyses showed an average change in the meta-analysis estimates of only 0.01 points. The worst case changed by 0.27 points (EF subscale in the small deterioration size class). The midpoints between estimates were calculated again for each scenario. The midpoints were generally robust to changes in the estimate used for the random effects variance. There were two midpoints altered by one point when using the upper confidence limit and one midpoint altered by one point if the lower confidence limit was used. The guidelines are therefore affected for the AP subscale (when using the lower limit) and the PA subscale (when using the upper limit) but the difference was only one point and affected at most two of the thresholds.

7.4 Summary of results and conclusions

It was possible to derive guidelines for the size of QOL differences from published mean differences using this methodology. Differences considered to be large by average expert opinion were rare and estimates of large differences were not possible directly from the meta-analysis, although approximate guidelines were still obtainable for some subscales.

For both cross-sectional and longitudinal comparisons, the guidelines for small and moderate differences varied according to subscale, indicating that a global rule across subscales is not appropriate. For a given scale and size-class, cross-sectional differences were different to the longitudinal differences, although not consistently smaller or larger. Longitudinal guidelines for changes for improvement differ from those for deterioration.

It should be noted that there was some overlap of confidence intervals between the estimates in adjacent size classes. This was potentially explained by small sample sizes in the longitudinal contrasts. Subscales were excluded from the guidelines if the meta-analysis estimates were not logical (i.e. medium estimates below the small estimates) and this could not be explained by a small numbers of contrasts. However, there was still some overlap between the 95% confidence intervals across the size classes.

Sensitivity analyses showed the method for imputing variances was acceptable and the impact of variation in estimates of around the random effects was minimal on the resulting guidelines.

The sensitivity analysis on the full dataset showed that not all contrasts were useful and appropriate to include in the meta-analysis. This analysis confirmed that the agreement criteria we devised had the desired effect of refining the contrasts to the better quality ones and as a result sensible guidelines were derived.

8 Patient interviews

8.1 Background

Although the main study was carried out using an expert panel we also wanted to develop a method for obtaining patient opinion on QOL differences. However, there were barriers to carrying out an identical task with patients.

Firstly, the EBIG project sought to quantify the size of differences between groups of patients rather than looking at changes for individual patients. The experts had to have clinical experience of similar groups of patients. In contrast, patients are not likely to have had experience of each clinical scenario, they are likely to have a fairly narrow experience compared to the wide range of clinical settings found in the papers.

Secondly, the experts also needed to be familiar with the QLQ-C30, so they could use their knowledge of the specific questions. Patients are also unlikely to have previously come across the QLQ-C30 questionnaire (or even the concept of measuring QOL).

Thirdly, the desired interpretation guidelines are partly aimed at obtaining sample sizes for clinical trials, and differences that result in a change in clinical practice are usually desired for this purpose. The definition of size class for the reviews was therefore based on 'clinical relevance' which would not necessarily be meaningful to patients.

Fourthly, the methodology required reviewers to read papers from medical journals which can be quite technical and likely to contain terminology unfamiliar to patients. There was also a practical issue with the number of comparisons and subscales a patient would be able to judge. Experts were involved in the study for up to three years and this would be a big commitment to expect from patients.

This pilot study was designed to explore if the methodology could be adapted in order to obtain patient reviews of the same data the experts were reviewing. In future studies, the methodology tested in this pilot study could enable expert and patient opinion to be combined for the development of the guidelines.

I conducted a further literature search to review how patients have previously been used to elicit opinions on the size of QOL differences. A number of other authors

have directly used patient opinion to elicit the minimally important difference in QOL scores from various instruments. Methods include using patients to rate each other(89;90) and within-patient global ratings(17;28;29;49;86). A number of these methods were discussed in Chapter 2. I did not find previous research where patients directly gave an opinion on QOL differences for groups of patients rather than individuals.

8.2 Ethical approval

This study was submitted to and received a favourable opinion from the Leeds East Research Ethics Committee and was approved by the local research and development (R&D) department. The full protocol is described in more detail in this chapter. The approved patient information sheet can be found in Appendix IV.

8.3 Objectives

8.3.1 Primary Objectives

- Can patients use information from published papers to form an opinion on meaningful differences in QOL scores?
- Can adequate familiarity with the QLQ-C30 and the way it produces quality of life scores be gained during an interview situation?

8.3.2 Secondary Objectives

- To what extent can patients form their opinion using data from a group of patients rather than their own individual experience?
- How do patients' opinions compare with clinicians opinions when using the same published data?
- Does the proposed interview need developing further prior to a larger study? In particular, is the information presented in a way patients can understand?
- Which types of scenarios should be developed for a further study?

8.4 Methods

The design and analysis of the study are reported as recommended in the consolidated criteria for reporting qualitative research (COREQ) checklist(91).

8.4.1 Study design

As this was a pilot study and it was largely unknown as to the best approach to getting patient opinions we felt an interview would be the best approach. This would allow plenty of discussion with the patients and opportunity to explore in-depth the way they approached the task and the reasoning behind their opinions. Cognitive interviewing(92) was used as this is aimed at gaining an understanding of how the patients approach their answers to questions in the interview.

I wrote an outline interview schedule and this was subjected to ethical review as part of the protocol. The schedule is displayed in Figure 50 and described more fully in the subsequent text.

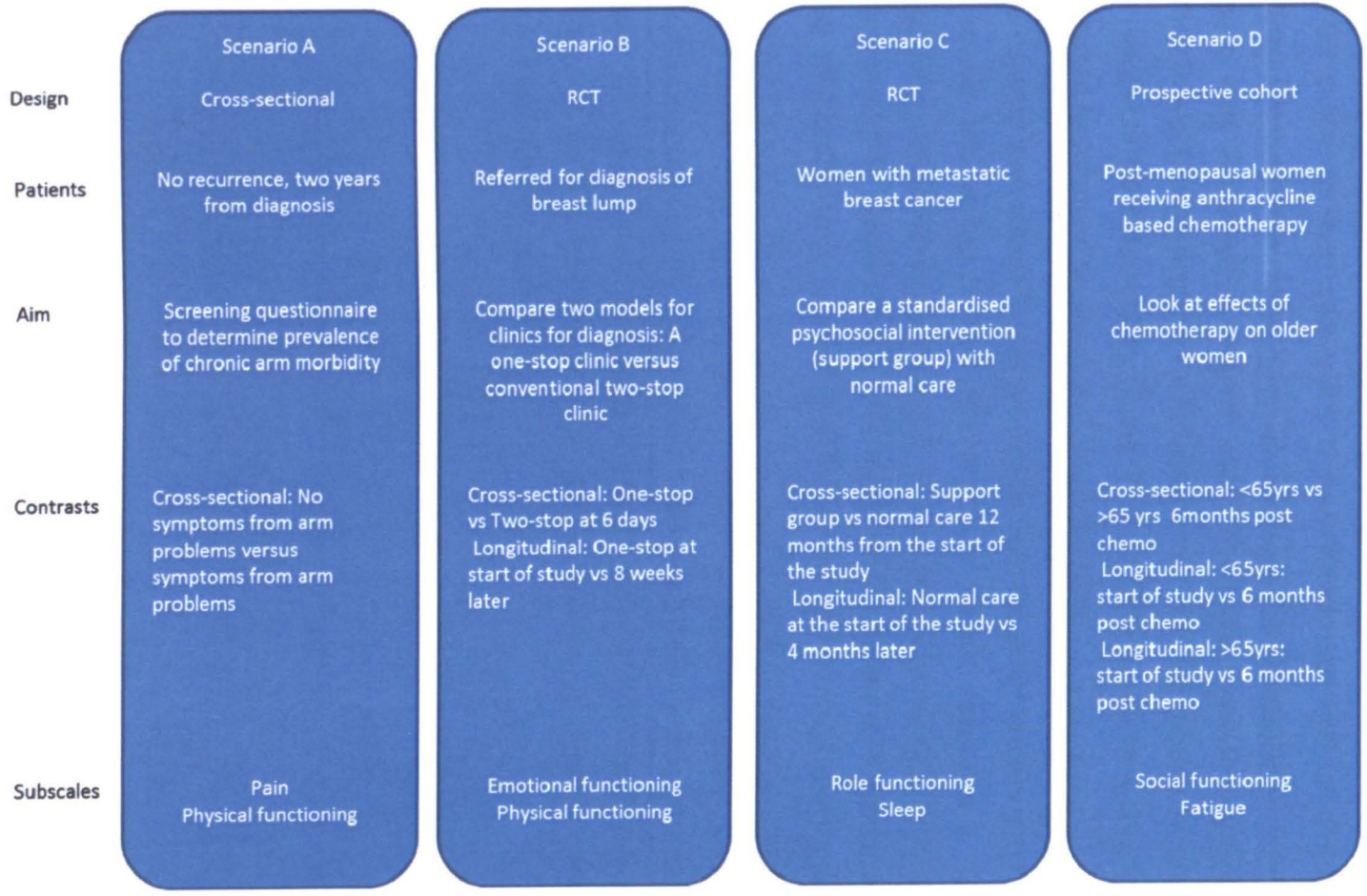
Figure 50 Outline interview schedule

1. Introduction to project and QLQ-C30 questionnaire
    a. Confirmation of agreeing to tape record interview
    b. Thank for volunteering and ask for questions at any time
    c. Explanation of project
    d. Introduction to questionnaire/scoring
    e. Explain problem of interpreting what scores mean to patients
    f. Explain that different scenarios will be presented and they will be asked for their opinion on the size of differences in QOL between groups of patients or patients over time
2. Show first scenario description; explain verbally, allow time to read and ask questions
3. Get patient opinions on two subscales for that scenario
    a. Highlight the questions involved in creating a score for that subscale
    b. Ask if there will be a noticeable difference between the groups and, if so, whether this difference would be small, medium or large
    c. Show pictures representing actual scores from these groups and ask if this changes their opinion
4. Show second scenario and repeat section 3
5. Explore what patient's needed to make a decision and their interpretation of small, medium and large.
    a. Were any of the scenarios easier to think about and why, do they think you need previous experience of the specific scenario in order to make a decision
    b. What did they think about to make their decision
    c. What did they think made a difference large, i.e. how were they interpreting the words 'large', 'medium' and 'small'
6. Interview close
    a. Questions or comments
    b. Feedback on the format of the interview and recommendations for any changes

The interview consisted of 4 parts; completion of informed consent, completion of socio-demographic details by the patient, an explanation of the purpose of the project and then the main content of the interview. The main interview started with familiarisation with the EORTC QLQ-C30 questionnaire and scoring. I showed the patients a copy of the questionnaire and described how each question contributed to a score for a subscale. I used an overlaying transparent sheet to demonstrate what the score would be when some of the questions were answered in a certain way and then a second overlay which showed how the score changed when one answer differed by only one category. Once the patient was happy with the questionnaire and the concept of scoring I moved on to discuss the first study scenario.

I developed four scenarios for use in the interview from published breast cancer studies also undergoing the expert review process. The studies chosen had all already been through the expert review process and had received three expert reviews at the time of designing this pilot study. I chose studies which demonstrated a range of disease stages, study designs, interventions and settings. I also looked at whether the experts agreed or disagreed on their assessment of the size of difference they would expect and included examples of both extremes. Finally I considered whether the actual QOL differences were large or small in order to include a range of actual sizes for the scenarios. I selected a variety of subscales across the scenarios but ensured some overlap between the subscales across the scenarios. Only two subscales were considered for each scenario to reduce the burden on the patient and length of the interview. Part of Scenario D is used as a worked example in Appendix V. A summary of the scenarios can be found in Figure 51.

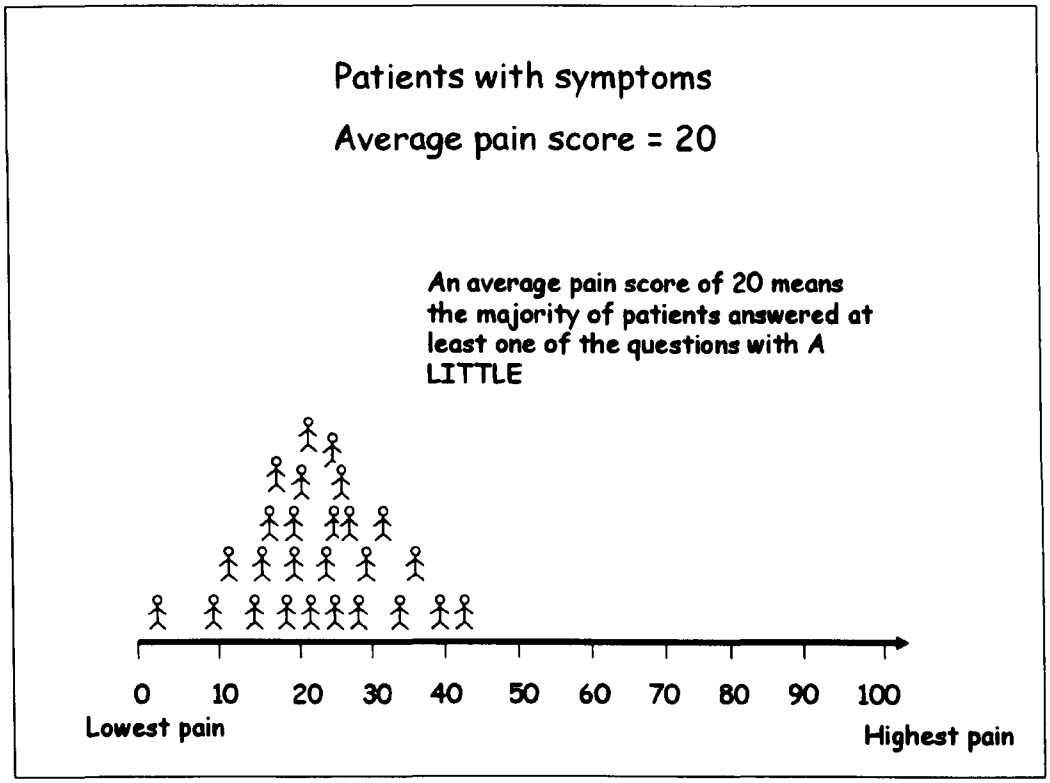
Figure 51 Patient interview scenario summary



During the interview I read a summary of the first study scenario with the patient and then asked if they wanted some time to read it again. The summary started with a description of the study and then described the two groups to be compared or one group and a description of the time points if it was a longitudinal contrast. I then described the subscales they were going to be asked to consider for that scenario, showing them the questions relevant to that subscale and examples of scores when the questions were answered in different ways. The patient was then asked to say whether they thought there would be a difference between the groups, or group over time, with respect to those subscales. If they thought there would be a difference they were then asked to comment on the size of that difference.

Patients were then shown the actual results from the study using a graph to show average scores and asked if that affected their earlier assessment. Graphical approaches have previously been shown to successfully convey QOL information to cancer patients(93;94). I developed graphs for this study based on a similar approach used in the PROMIS® study, *www.nihpromis.org(95)*. The format I used showed the actual means for each contrast using text and 'stick people' to represent the group of patients on a graph. They were designed to show that the mean score comes from a group of individuals with varying QOL scores. I showed patients a graph for each of the two groups in the contrasts so they could use them to re-consider their opinion on the size of difference. An example of one of the graphs is shown in Figure 52 and further examples can be found in Appendix V.

Figure 52 Example graph to show patients the actual QOL means



Once the patients had looked at the actual scores and commented on how it may change their original assessment of the QOL difference I then showed them the results of the expert review for the same contrasts and probed for their comments.

Finally, there were some more general questions at the end of the interview designed to find out if particular subscales or scenarios worked better than others and to find out how the patient had approached the task.

8.4.2 Interviewer training and pilot testing

Since I had no previous experience of interviewing patients I received some training from the CRUK Psychosocial Oncology Group at St James' University Hospital. Once I had written the interview I piloted the format on one of the research assistants who had experience of conducting patient interviews. I also shadowed a research assistant in the breast cancer clinic to observe the patient consent process before I approached patients for this study.

8.4.3 Patient sample and recruitment

There are no set guidelines for the appropriate size of sample in qualitative research. Generally qualitative studies require much smaller numbers of patients than quantitative studies. Using the principle of saturation, patients are recruited until a point

is reached where no new information emerges from subsequent interviews. As the interviews here were focussed on specific issues (in-depth rather than broad-ranging) and the content of the interview was novel and experimental, I considered that a maximum of 12 patients should be adequate to achieve the study objectives.

Purposive sampling was used to obtain a heterogeneous sample with respect to age and extent of disease according to the matrix in Table 61. These two factors were chosen as the most appropriate ones in order to obtain a sample containing a broad range of breast cancer patients. Recruiting up to 12 patients allowed for one to two patients per group and provided adequate opportunity to interview patients across a broad range of ages and with varying degrees of disease.

Table 61 Purposive sampling matrix

Extent of disease	Primary local		Local recurrent		Metastases		Disease-free	
Age	<50	≥50	<50	≥50	<50	≥50	<50	≥50

Patients were eligible for inclusion if they met the following inclusion criteria and none of the exclusion criteria applied:-

Inclusion criteria:-

- Ability to read and understand English
- Patients with a diagnosis of breast cancer
- Patients more than 18 years of age
- Patients who have received at least three months of treatment for their breast cancer

Exclusion criteria:-

- Patients who have just been informed about diagnosis of primary or recurrent disease
- Patients already participating in a psychosocial oncology research projects
- Patients attending their first oncology appointment

Potentially eligible patients attending breast cancer clinics were identified using the Patient Pathways Manager (PPM) system. Patients were identified by a Research Assistant member of the CRUK Psychosocial Oncology Group.

I approached eligible patients consecutively (in order of arrival) in the waiting room of a breast cancer clinic at St James University Hospital. I asked if they were

willing to talk with me about participation in a research study and if so I then briefly described the study to them while they were waiting for their appointment. If the patient was still interested I gave them patient information sheet and consent form to take away so they could read it in detail and decide whether to participate. Patients gave their phone number if they wished to be contacted regarding possible participation and to arrange an interview date. All patients had at least 24 hours to decide if they wished to participate.

8.4.4 Interview content and setting

I used two of the study scenarios during each interview. I wanted to use all four scenarios in the first two interviews in order to test them in case of any problems with the design of the scenarios or interview script. Following these first two interviews I wanted to use the scenarios equally where possible and ensuring that a range of patients received each scenario. I also ensured that the scenarios were used in different orders during the interviews, in case the patient was more rushed or fatigued by the second scenario. The first two interviewees received scenarios A then B and scenarios C then D respectively. Once the four scenarios had been tested in these first two interviews I chose scenarios for each interview according to the characteristics of the next patient in order to achieve the desired mix of ordering and patient characteristics using each scenario.

I interviewed patients at St James University hospital (if they were attending for an appointment in the few weeks following agreeing to take part in the study) or at their home if they preferred. The interviews were audio-recorded so that they could be transcribed in full for the analysis. This allowed for some external validity checking of the analysis (as detailed in section 8.4.6).

8.4.5 Analysis methods

8.4.5.1 Transcription

The recorded interviews were transcribed verbatim as soon as possible following the interview. I transcribed the first interview and then used a third party to complete the subsequent transcriptions. Non-relevant discussion was excluded from the transcripts (e.g. telephone calls received by the patient during interview, use of the word 'umm' etc). Place names or names of people were anonymised during transcription. Interview recordings were identified only by study number for the purposes of confidentiality during transcription. I verified all of the transcriptions by listening to the full interview and comparing against the transcript.

8.4.5.2 Theoretical framework

Framework analysis(96) was used to analyse the interview transcripts. I chose this as it closely follows the information from the interviews and reports the data fully. It was suitable for this study as I had defined set objectives in the protocol which would guide the information required from the interviews. These objectives could therefore form the basis of the framework. Framework analysis allows for development of the framework during the analysis and since this methodology was being tried for the first time this seemed appropriate as it also allowed for information to emerge that had not been considered *a priori*. No specific software for analysing qualitative data was used for the analysis. I coded transcripts by hand.

Framework analysis was undertaken in stages as described by Bryman and Burgess(96):-

- Stage 1: Familiarisation

I prepared initial summary sheets immediately after each interview containing any relevant notes regarding the interview. Verifying the transcripts was also part of the familiarisation process, and I usually carried this out after every three interviews. On verifying the transcription I expanded the summary sheet by adding any key ideas or themes I saw emerging. I also included a summary of the patients' responses to the size of differences for each contrast so these were more easy to compare later across patients and with the experts.

- Stage 2: Identifying a thematic framework

The protocol was used to identify an initial thematic framework by using the objectives and other specific questions the study aimed to address. There were 11 items in the initial framework, the shortened descriptions in bold were used as the theme headings:-

1. **Ability to use published data**: Can patients use information from published papers to form an opinion on meaningful differences in QOL scores?
2. **Familiarity with questionnaire and scoring**: Can adequate familiarity with the QLQ-C30 and the way it produces quality of life scores be gained during an interview situation?
3. **Group vs individual**: To what extent can patients form their opinion using data from a group of patients rather than their own individual experience?

4. **Patient vs clinician**: How do patients' opinions compare with clinicians opinions when using the same published data?
5. **Interview development**: Does the proposed interview need developing further prior to a larger study? In particular, is the information presented in a way patients can understand?
6. **Scenario feedback**: Which types of scenarios should be developed for a further study? Do patients find some scenarios easier to think about and why?
7. **Understanding of questionnaire and scoring**: Do patients understand the questionnaire and scoring system given the description and examples in the interview?
8. **Information used to inform decision**: An understanding of what information patients use to decide if groups of patients will have a noticeably different QOL from each other.
9. **Similar experience to scenario required**: Do patients need to have a similar personal experience to identify with groups of patients?
10. **Definitions of large, medium and small**: An understanding of the way patients define large, medium and small differences.
11. **Impact of knowledge of actual scores**: How do patients' opinions differ if they have knowledge of the actual scores compared to when they do not?

I developed the initial thematic framework further during the analysis using three interviews at a time. The initial framework was coded with items 1 to 11 as above and I used these codes in the margin of the transcripts to highlight relevant sections of interview script. Any new or emergent themes were added to the framework in the order they were found in the transcripts.

- Stage 3: Indexing

Once I had reviewed all of the interviews, the codelist was finalised. The final codelist was then systematically applied to each transcript by reading the transcripts again and annotating the whole transcript with codes where the text was relevant to any of the themes in the final codelist. Numerical codes were listed in the margin for each section of text as applicable, with additional notes if I felt these would be useful.

- Stage 4: Charting

I drew a chart for each theme identified in the final codelist. This was in the form of a table that divided the broad themes into sub-themes and included quotes from patients to show examples of the theme. The overall theme was used as the chart heading. Sub-codes within the theme were used as columns and each patient had a row in the chart. I maintained the same order (interview order) in each chart so that individuals could be reviewed across the themes where required. Charting in this way was used to group excerpts from the interview text. Sometimes the text was copied and pasted verbatim into the chart and sometimes a distilled summary was used in the charts. However, I always referenced the original text so the source could be traced. I referenced text using a code consisting of the interview number, scenario (where the text was specific to a scenario) and page number.

- Stage 5 Mapping and interpretation

Once the charts were completed I reviewed them to look for links between them and then regrouped any related codes together. The main themes were the objectives from the protocol and under each theme either whole charts or certain sub-themes from charts were combined during this process. I used major themes to define an emergent or recurrent theme relating to the overall objective. Minor themes were then used to categorise the range of responses from patients within those major themes. I used direct quotes to illustrate the themes where applicable. These final tables of major and minor themes along with supportive excerpts from the interview texts are used later to summarise the results from the study in section 8.5. The full charts are reported in order to show as much detail behind the analysis as possible, however, I have also provided a summary of the results for each objective and then overall conclusions at the end of the chapter.

I analysed the transcripts and conducted further interviews iteratively. This is recommended in qualitative research to ensure that saturation is achieved, that is no new or emergent themes are still being found when the sample size has been reached.

8.4.5.3 Quantitative analysis

I also used various quantitative summaries to supplement the qualitative analysis. In order to compare the expert and patient opinion, the patients' opinions were put on the same scale to the experts. The words 'small', 'medium' or 'large' used during the interview were converted to 1, 2, or 3 respectively. Zero was used where patients did not think there would be a difference. Where patients described differences in between categories (e.g. small to medium), an increment of 0.5 was used. Positive or negative

scores were used to indicate the direction of the difference. Because of the way the expert scores were derived (using weighted averages) these were on a more continuous scale than the patient data.

Concordance between patients reviewing the same contrasts was used to inform whether patients have the ability to use the published data in this way. I used a similar method to that used for the experts, looking at the maximum distance (in terms of the number of size classes) between the patient scores. This was aimed at giving an indication as to whether using a panel of patients may be feasible in the same way as the expert panel was utilised. Scatter graphs and correlation coefficients were used to look at how the patient opinion compared with the actual scores from the studies, in the same way as I did for the expert scores. I also repeated this for the expert scores but only using the subset of the contrasts used for these patient interviews rather than the full dataset. I used bar charts to illustrate how patient and expert opinions differed on average for each contrast.

8.4.6 Validation of analysis

I carried out all of the interviews and analysis so in order to check the results for bias and ensure the validity of the conclusions from the study, the rest of the study team (Prof Velikova, Prof Fayers, Prof King and Prof Brown) each reviewed one transcript and checked the coding according to the final code-list. Any disagreements in coding or new themes were discussed as a team and an agreement reached.

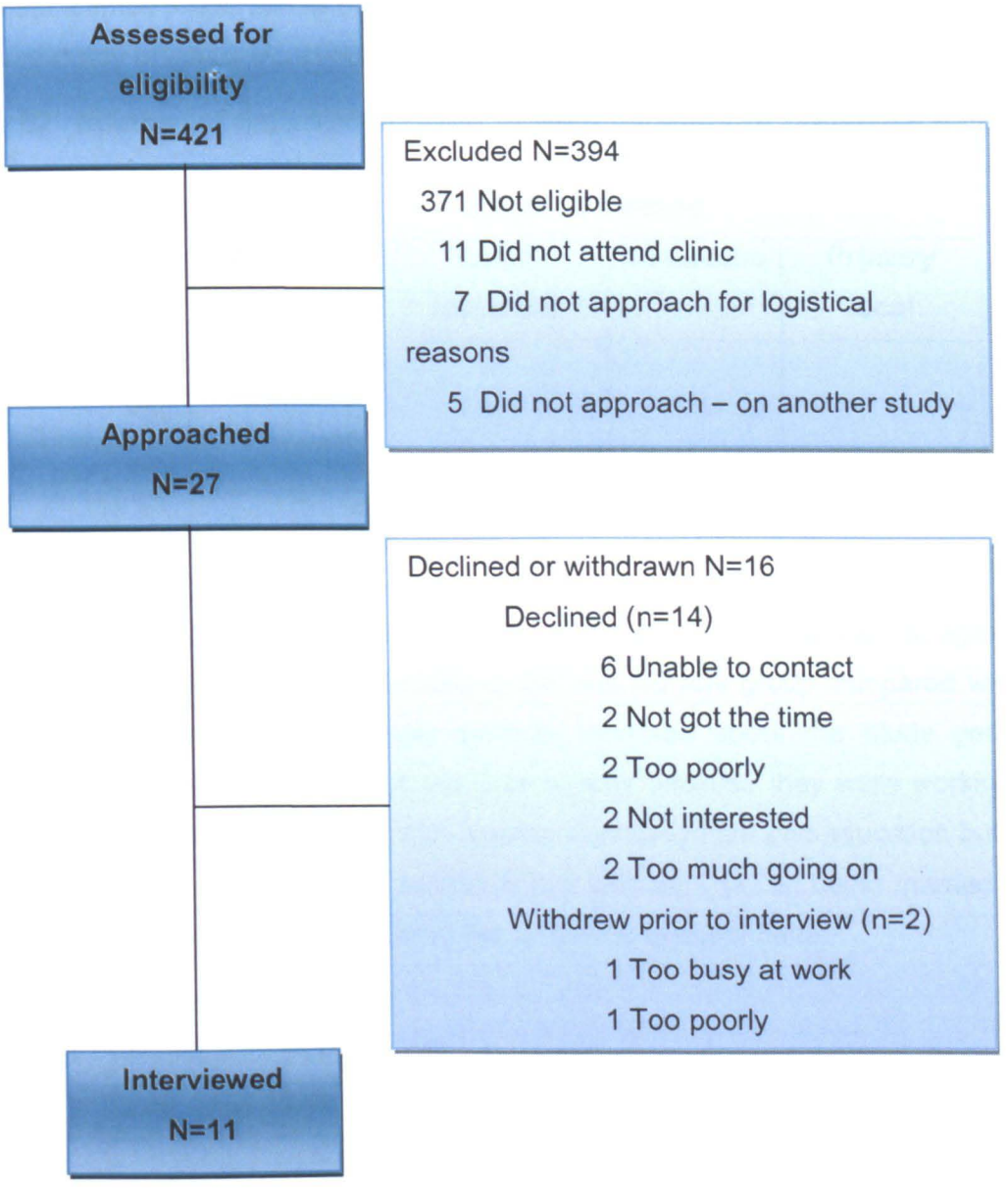
8.5 Results

8.5.1 Sample

I recruited patients at St James' University Hospital in Leeds over 13 weeks, from 28/7/2008 to 27/10/2008. Figure 53 shows details of the numbers of patients approached and recruited for the study. 421 patients were assessed for eligibility and 50 potentially eligible patients were identified. I approached 27 of these and 14 declined. Two patients agreed to take part but subsequently withdrew prior to the interview. Not all patients were approached for practical reasons. Firstly, I could not attend all of the clinics where patients were presenting and secondly because the transcription and analysis was taking place iteratively I needed to allow time in between interviews to code the previous transcripts.

I conducted 11 interviews, with six taking place at patients' homes, four in a hospital interview room and one in the hospital out-patient treatment ward. The interviews lasted between one and two hours.

Figure 53 Flow of patients through qualitative interview study



Study patients had a range of ages and disease extent, as planned by the purposive sampling (Table 62). However, I did not recruit any patients in one of the categories (local recurrent less than 50 years old). On discussion with Prof Velikova we identified two possible reasons for this. Firstly, younger patients with recurrence represent a small proportion of breast cancer patients and secondly these patients may not necessarily attend the particular clinic where recruitment was carried out.

Table 62 Number of patients recruited in each category of purposive sampling matrix

Number of patients	Extent of disease				Total
	Disease-free	Local recurrent	Metastatic	Primary local	
Age					
<50	1	0	2	1	4
≥50	2	1	2	2	7
Total	3	1	4	3	11

Participating patients ranged from 33 to 78 years old, with a median age of 57 (Table 63). More patients were recruited in the over 50 age group compared with the under 50s (64% vs 36%). Younger patients informed about the study generally declined due to time available, and this was usually because they were working full-time. Patient characteristics varied with respect to employment and education but were fairly homogenous with respect to marital status with 9/11 (82%) being married. Only one patient had previously encountered the QLQ-C30 questionnaire.

Table 63 Patient characteristics

Number of patients (%)	
Age (yrs): Median (range)	57 (33 to 78)
Extent of disease	
Disease-free	3 (27.3%)
Local recurrent	1 (9.1%)
Metastatic	4 (36.4%)
Primary local	3 (27.3%)
Marital status	

Number of patients (%)	
Married	9 (81.8%)
Separated/Divorced	1 (9.1%)
Single	1 (9.1%)
Number of occupants in house	
1	1 (9.1)
2	6 (54.5%)
3	2 (18.2%)
4	2 (18.2%)
Current employment	
At home	1 (9.1%)
Retired	4 (36.4%)
Unemployed	1 (9.1%)
Working full time	3 (27.3%)
Working part time	2 (18.2%)
Educated after minimum school leaving age	8 (72.7%)
Degree/professional qualification	5 (45.5%)
Previous experience of QLQ-C30	1 (9.1%)

Around one third of patients had metastatic breast cancer (Table 64). Two-thirds of patients were currently receiving chemotherapy. Most patients (82%) had received previous lines of treatment.

Table 64 Details of cancer type and treatment

	Number of patients	%
Extent		
Metastatic	4	36.4
Disease-free	3	27.3
Primary local	3	27.3
Local recurrent	1	9.1

Diagnosis		
Infiltrating ductal carcinoma	7	63.6
Metastatic	2	18.2
Carcinoma - unspecified	2	18.2
Stage at diagnosis		
1	1	9.1
2	2	18.2
3	3	27.3
4	4	36.4
Unknown	1	9.1
Current treatment		
Chemotherapy	7	63.6
Hormonotherapy	2	18.2
Other	1	9.1
Follow up	1	9.1
Previous treatment		
Chemotherapy/Radiotherapy/Surgery	2	18.2
Chemotherapy/Radiotherapy/Surgery/Hormonal therapy	2	18.2
None	2	18.2
Hormonal therapy	1	9.1
Radiotherapy/Bisphosphonates	1	9.1
Surgery	1	9.1
Radiotherapy/Surgery/Hormonal therapy	1	9.1
Chemotherapy/Surgery	1	9.1

8.5.2 Scenario allocation

Each scenario was used five or six times during the study (Table 65). The scenarios were allocated as described in the methods, with attention to the ordering and ensuring that all scenarios were encountered between the patients in a disease and age group (where numbers allowed). A total of 73 opinions on the contrasts were collected.

Table 65 Scenario allocation and contrasts reviewed

Interview number	Disease group	Age group	First	Second	Contrasts (Type:Subscale)*	Total number of contrasts reviewed
1	Mets	<50	A	B	X:PA X:PF X:EF X:PF L:EF L:PF	6
2	DF	>50	C	D	L:RF L:SL X:RF X:SL L1:FA L2:FA X:SF	7
3	Primary	>50	B	D	X:EF X:PF L:EF L:PF L1:FA L2:FA X:SF	7
4	Mets	>50	A	C	X:PA X:PF L:RF L:SL X:RF X:SL	6
5	DF	>50	A	B	X:PA X:PF X:EF X:PF L:EF L:PF	6
6	Primary	<50	C	D	L:RF L:SL X:RF X:SL L1:FA L2:FA X:SF	7
7	Mets	>50	B	D	X:EF X:PF L:EF L:PF L1:FA L2:FA X:SF	7
8	Mets	<50	D	C	L1:FA L2:FA X:SF L:RF L:SL X:RF X:SL	7
9	Local rec	>50	C	A	L:RF L:SL X:RF X:SL X:PA X:PF	6
10	Primary	>50	B	A	X:EF X:PF L:EF L:PF X:PA X:PF	6
11	DF	<50	C	B	L:RF L:SL X:RF X:SL X:EF X:PF L:EF L:PF	8
				Total contrasts:		73

*Type X=Cross-sectional, L=Longitudinal

8.5.3 Development of thematic framework

During the review of transcripts three new themes were added to the initial thematic framework; approach to the task, general feedback and personality type. These were themes that added information regarding how patients were approaching the task and why they may approach it in a particular way. These themes were

subsequently combined across the other data themes and therefore do not appear as major themes in the final framework. For example, the approach to the task theme was merged with the 'Group vs Individual' theme as I noticed the content was related during the mapping and interpretation stage of the analysis. As any new themes had been re-coded in with the original objectives and no new themes were still being identified in the last set of interviews, it is probable that saturation in terms of themes had been reached during the 11 interviews.

The final thematic framework is summarised in Table 66. The overall themes represent the study objectives. The major and minor sub-themes were used to group the excerpts from the transcripts, further categorising the content of the interviews coded under the overall theme. The purpose of this table is to show the final framework used to code all of the transcripts. Further detail on each theme, including quotes from patients, are reported in the sections that follow.

Table 66 Final thematic framework

Theme	Sub-themes	
	Major	**Minor**
Ability to use published information	Confidence in their answers	Confidence comes with experience
		Lack of confidence
		Guessing
	Showing in-depth understanding of QOL concepts	Discrete nature of scoring
		Response shift/recall bias
		Detailed description of QOL concept
		Questions in subscale may go in opposite directions and cancel out an overall difference
		Groups in scenario may have a different reference point
	Confidence in project	Need a mix of patients
	Hard without experience	Hard for a layman
		Different types of cancer/chemo have different effects
		More accurate with experience
		People are so different
		View may be different with experience

Theme	Sub-themes	
	Major	**Minor**
	Experience not necessary	Can use own experience and others
		Experience does not have to be identical
		Can think along same lines
		Subjective rather than objective
		Can empathise
		Not for more general scenarios
	Difficult even with experience	Scenario needs more info
		Response shift/recall bias
	Judgements based on experience/belief	Strong belief/preference driving judgements
		Own experience driving judgements
Familiarity with questionnaire and scoring	Familiarity	Reference to subscale questions while answering
		Linking subscales
		Putting self onto scales
	Understanding of questionnaire	Querying or commenting on the timing of questionnaire or period covered
		Awareness of multidimensional nature of QOL
		Showing an understanding of how QLQ-C30 may be useful
	Understanding of scoring	Confusion/clarity with direction of scoring indicating improvement/deterioration
		Reference to scoring system when thinking about answers
		Understanding that your baseline score is important
		Suggesting specific score/category for group
	Showing an understanding of the graphs	Confirming answers vs graphs
		Interpreting before explanation given
Group or individual thinking	Group	Awareness of variation in group
		Clear group thinking
		Reference to others in the abstract
		Reference to experience of others

Theme	Sub-themes	
	Major	**Minor**
		Thinking of majority
		Finding average difficult
	Individual	Reference to self in the abstract
		Reference to own experience
		Answering subscale for self
	Approach to task	Using own experience then generalising
		No reference to own experience
Patient vs clinician opinions	Ability of clinicians to judge	Able
		Not able
	Reasoning for differences	Can use their experience
		Experts may assume positive outcome
		Place different importance on QOL changes
		Can see reasons why differences could be either way round
	Reasoning for agreement	
Definitions of small, medium and large differences	Specific definition of small	Treatment working; no pain and able to function
	Specific definition of medium	Sizeable
		Ups and downs
	Specific definition of large	Extremes
		Own experience vs others
		A little vs A lot
		A marked difference
		Significant difference on your life
	Using questionnaire responses (8)	A little, quite a bit, very much
	Other wording	Percentages
		Marked
		Slight
		Average
		Substantial
	Wording irrelevant	

Theme	Sub-themes	
	Major	**Minor**
	Not specific	Gut instinct
Presentation and understanding	Confusion with details of scenario	Forgetting timelines of patients from diagnosis
		Patient referring to pain but scenario is regarding symptoms
		Patient referring to normal care group as having no support
	Confusion with required task	Needing clarification of difference rather than by group
		Trying to complete task with reference to self rather than scenario
		Not understanding purpose of scoring explanation
	Graphs	Own interpretation
		Confusion with direction of difference
		Influenced by individuals/extremes on graph
		Useful
	Understanding of scenario	Linking subscales and using previous subscale answer
		Awareness of timescales in scenario and impact
		Showing in-depth understanding of groups in the scenario
		Showing in-depth thought around subscale
Scenario development	Identifying difficult scenarios	Due to variation in group or differing effects on individuals
		Due to lack of experience
		Due to their experience not being the norm

8.5.4 Primary objective (1): Can patients use information from published papers to form an opinion on meaningful differences in QOL scores?

The primary objective was addressed by considering both quantitative and qualitative data from the interviews. I considered how many contrasts out of the total given were regarded as impossible to judge by the patients and looked at the concordance between patients judging the same contrasts. The charting stage from the qualitative analysis provides insight into whether patients felt able to use the published information.

8.5.4.1 Individual contrasts patients felt unable to judge

There were only six instances (occurring in three interviews) where patients felt unable to form an opinion. There were a total of 73 judgements on contrasts in the interviews (see Table 65), therefore patients felt able to form an opinion in 92% of them. Five of the identified issues were from scenario C and one was from scenario D. Three of the contrasts were cross-sectional and three were longitudinal. Affected subscales were RF, SF and SL. Reasons for not being able to form an opinion were due to a perceived lack of information in the scenario (four contrasts) or feeling that the effect of the intervention being studied would be very individual, benefiting some but having a negative effect on others, meaning that the group average could go either way.

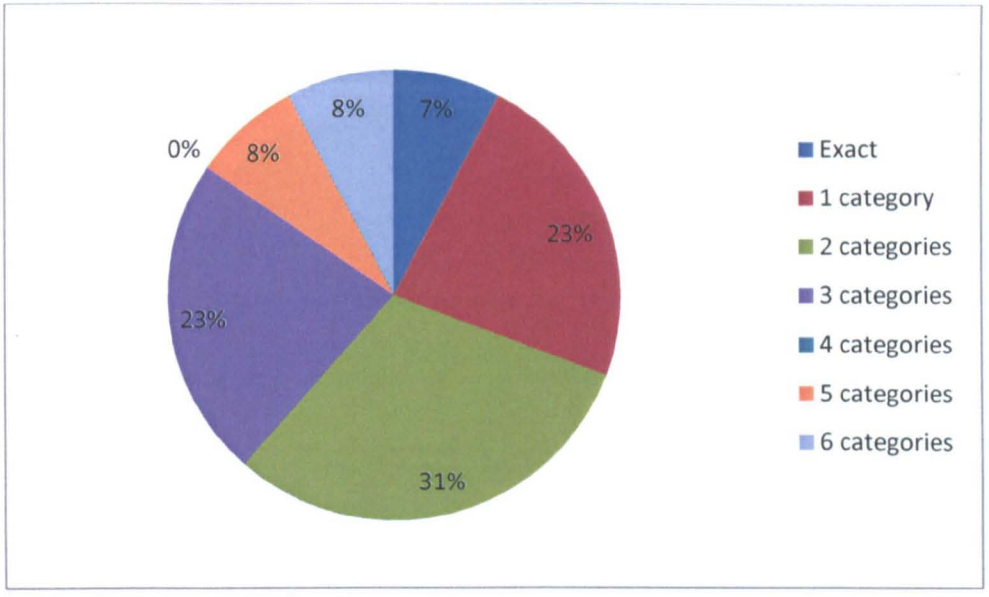
In order to try to address the lack of information in the scenarios for the remaining interviews extra information about the study was added to these scenarios. Treatment information was added to scenario C and an average age and range for patients was added to scenario D. No further problems emerged in later interviews therefore it seems likely that this is a shortfall of the original scenarios I designed rather than reflecting the patients' inability to use published data.

8.5.4.2 Concordance between patients – blinded to actual scores

There were 13 different contrasts represented in the four scenarios and five or six patients reviewed each of the contrasts (see Table 65). Perfect agreement between patients only occurred for one of the thirteen contrasts (where all patients deemed there would be no difference). 61% of the contrasts had up to a maximum difference between patients of two size classes (Figure 54). 39% had maximum differences of three or more size classes. Although scenario B contained the contrast with perfect

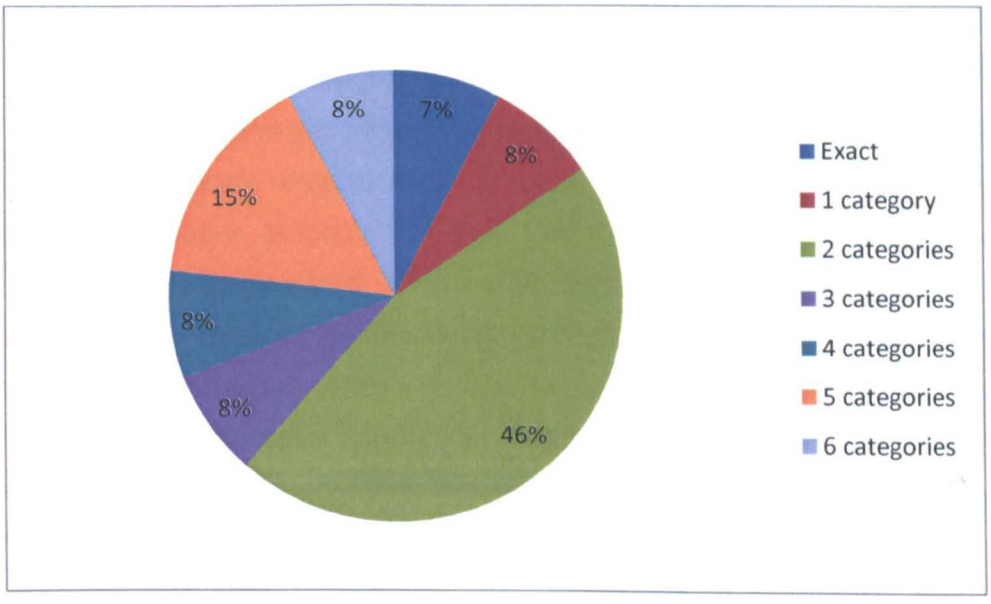
agreement from patients the other three contrasts in that scenario resulted in a wide range of expected differences (from none to large). Scenario D resulted in the most disparity regarding the direction of the expected scores.

Figure 54 Distance between patients' reviews (blinded to actual scores)



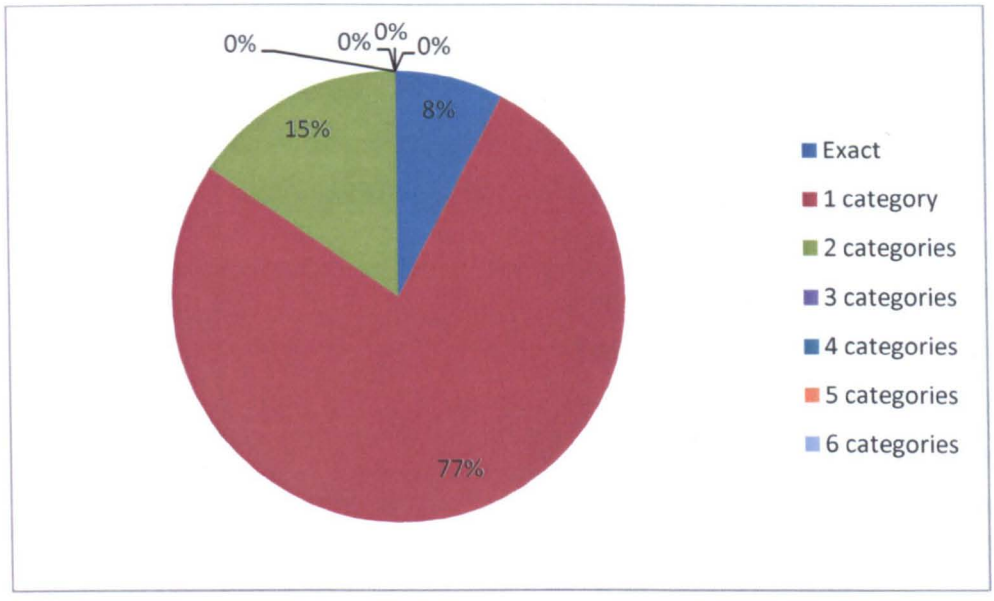
After seeing the actual QOL scores, complete agreement between patients still only occurred for the same one contrast. There was a slight improvement in agreement by showing patients the actual scores, with 62% of contrasts where patient reviews were up to two size classes apart (Figure 55).

Figure 55 Distance between patients' reviews (after seeing actual scores)



For this subset the experts had 85% of the contrasts where they disagreed by a maximum of one size class and all contrasts had the experts with a maximum of two size classes apart (Figure 56).

Figure 56 Distance between the expert scores for the subset of contrasts undergoing patient review



8.5.4.3 Qualitative results

Seven major themes were extracted from the data which were felt to inform on the question of patients' ability to use the published information (Table 67). These were labelled as follows; confidence in their answers, showing in-depth understanding of QOL concepts, confidence in project, hard without experience, experience not necessary, hard even with experience and judgements based on experience/belief.

While patients generally provided their expectation of the QOL difference, there were often fairly flippant remarks about whether they'd 'get it right' before seeing the scores from the paper, particularly if they had already seen a result they were not expecting in a previous contrast. Whilst a number of patients commented on the fact that it was hard to think of your expectation of the scores without experience of the situation in the study, they also commented on how they still felt able to make judgements without the direct experience. There was a feeling that they could relate to the scenario if they had experienced something similar or could draw on others experiences as well as their own but that with experience their answers may differ. It was apparent that often a patient's own experience or preference was a factor influencing their expectations of QOL differences, particularly if they had strong feelings regarding the situation in the scenario. Patients demonstrated some in-depth thinking

around QOL concepts, particularly thinking through different aspects of the concept being measured.

Table 67 Ability to use published information

Major themes (number of interviews where extracted)	**Minor themes (number of interviews where extracted)**	**Quotes**
Confidence in their answers (7)	Confidence comes with experience (1)	But you certainly feel more confident answering things that you have a slight knowledge about as to how it's affected you, albeit I know it obviously affects different people differently, some people you know, I think have been more affected than me and other people you know, much less really. (6 p23)
	Lack of confidence (6)	I didn't really get the questions right (8 p18) I'll be totally wrong here (5 B p21)
	Guessing (1)	
Showing in-depth understanding of QOL concepts (6)	Discrete nature of scoring (1)	Because there's only four options I think the movement between them is quite hefty (11 C p5)

Major themes (number of interviews where extracted)	**Minor themes (number of interviews where extracted)**	**Quotes**
	Response shift/recall bias (1)	*<Talking about the support group>* ...may have answered more realistically or honestly because they were told to perhaps look at how they're feeling (6 C p16) People near the end of the treatment may have more idea...but then they might have forgotten how it was (6 D p11)
	Detailed description of QOL concept (3)	Pain can be different cause it can be depending how...how not always necessarily the severity of it but the frequency of it if you've got it sort of permanent sort of pain there then it can affect how you feel on the day to day basis and it's a nagging thing and it drags you down sometimes (4 C p36) Don't know what the doctor would class it as, like feeling down but not clinical depression (1 B p13) There's a physical and a mental tiredness (3 D p21)

Major themes (number of interviews where extracted)	**Minor themes (number of interviews where extracted)**	**Quotes**
	Questions in subscale may go in opposite directions and cancel out an overall difference (1)	
	Groups in scenario may have a different reference point (1)	I would say it was quite a big difference because I think even if, because I'm trying to think whether someone would have said I've noticed quite a bit because they're comparing themselves to, if they're not in the support group they're comparing themselves to people who they are surrounded by everyday who are kind of 100% able to do everything they want to do so you'd feel, emotionally you'd feel that you couldn't do as much as them whereas if you were in your support group you'd see that everyone else has similar problems so it'd make you feel that you were more, not normal but more average. (11 C p5)

Major themes (number of interviews where extracted)	**Minor themes (number of interviews where extracted)**	**Quotes**
Confidence in project (4)	Need a mix of patients (4)	I think you'll crack it but I think it might, it's weighing up this business of the groups, but with it being an individual thing...I could tell you all sorts of things...but the person here at the side of me that's probably my same age might not think the same (9 C p21) I do think in the end the more people that you ask, do this test for you know, I think everybody's going to have different opinions and answers and I think when its all put in the mix it will come out right won't it (5 p29)

Major themes (number of interviews where extracted)	**Minor themes (number of interviews where extracted)**	**Quotes**
Hard without experience (8)	Hard for a layman (5)	Not knowing and having them symptoms it's very hard for a layman to judge how much pain they would be in (1 A p6) Having never experienced that I've been lucky I didn't have any arm pain or anything so it's hard to know in affect just how it would feel, how severe it might be (4 A p15) If you've not had it at all then you will not understand...I think it's easier if you're going through the same process (8 p17)
	Different types of cancer/chemo have different effects (1)	if you've got hormonal cancer like me then yeah its totally different you become a weeping wailing wreck ...some having different chemos their emotions are different (1 p25)
	More accurate with experience (3)	
	People are so different (1)	even though I had breast cancer I can't put myself in somebody else's situation and...I mean people are just so different aren't they? (5 A p13)

Major themes (number of interviews where extracted)	**Minor themes (number of interviews where extracted)**	**Quotes**
	View may be different with experience (2)	I was answering stuff I haven't really got a clue about, I only answered on what I saw and you know sometimes you see things and you make a totally different judgement don't you (1 A p29) Perhaps if I had been involved in something like that I might have a different view on it, I don't know (6 C p16)
Experience not necessary (9)	Can use own experience and others (1)	And people you know that you've met in the clinic what's happened to them (2 p27)
	Experience does not have to be identical (3)	I mean I still get a bit of pain in my arm where the chemo strips your veins, you know that stops you from doing things (5 A p6) I'd experienced some other thing that had restricted me movement in a ... so I could kind of think of that... and I, just the difference it made to my life in a sense for a few weeks till I had some treatment for it (4 A p43)
	Can think along same lines (2)	

Major themes (number of interviews where extracted)	**Minor themes (number of interviews where extracted)**	**Quotes**
	Subjective rather than objective (1)	I suppose it's subjective in a sense, it's not as objective as other people looking at it but it's still, it's yeah, you've got something to base your opinion on in a sense (4 A p18)
	Can empathise (1)	
	Not for more general scenarios (1)	your clinics yeah you can ask any cancer patient that because we've all been there in some clinic (1 A p28)
Difficult even with experience (2)	Scenario needs more info (1)	I've got that and I'm in the same situation then, in a sense...but just given on the bare bones like this then it would be quite difficult (4 C p33)
	Response shift/recall bias (1)	People near the end of the treatment may have more idea...but then they might have forgotten how it was (6 D p11)
Judgements based on experience/belief (8)	Strong belief/preference driving judgements (7)	I do honestly believe that them support groups as you go further on in your cancers (9 p32) I can only go on how I would feel and I say I am not one that overly stresses but I'd prefer to know straight away (3 B p25)

Major themes (number of interviews where extracted)	Minor themes (number of interviews where extracted)	Quotes
	Own experience driving judgements (3)	I was fine and when other people were being sick and feeling really really tired, I think that's a big difference and I think if I'd have been like that my answers would have been different to what I've given (5 p27)

8.5.4.4 Summary: Can patients use information from published papers to form an opinion on meaningful differences in QOL scores?

There were very few instances where patients felt they could not give an opinion on a contrast. Generally it appeared that patients felt able to form an opinion using published data on meaningful differences in QOL scores as long as there was sufficient information in the summary of the study provided in the scenario.

The distance between reviews from different patients was quite high. The reviews were more than three categories apart for nearly half of the contrasts reviewed. This was only slightly improved when patients were asked to judge again after seeing the actual QOL results. This is compared to the expert reviews where less than 10% of the reviews had a distance of three or more categories between the experts.

It was apparent that although patients were judging the expected QOL differences between groups of patients, the patient's own experience and values influence their opinions. This is particularly noticeable where patients had an especially positive or negative treatment/disease experience.

8.5.5 Primary objective (2): Can adequate familiarity with the QLQ-C30 and the way it produces quality of life scores be gained during an interview situation?

Only one patient definitely had experience of filling in the QLQ-C30 questionnaire before (as evidenced by participation in a previous study) but three other patients thought it looked familiar or thought they had filled in similar questionnaires before. Even the patient who was identified as having previously filled in the questionnaire was

vague about whether she recalled the questionnaire therefore essentially all patients were starting from a similar point in terms of lack of familiarity with the questionnaire prior to the interview.

8.5.5.1 Qualitative results

A few patients asked questions during the description of the QLQ-C30 at the start of the interview and some commented on the fact that symptom and functioning scales were scored in opposite directions. Generally patients gave the impression they understood the explanation of the questionnaire.

Four major themes were extracted from the interviews relating to evidence of familiarity with the questionnaire gained during the interview or understanding of the questionnaire/scoring system (Table 68).

Some patients referred back to the individual questions making up the subscale while answering, showing they were specifically thinking of how the questionnaire was measuring the QOL concept. Some referred to the QOL concept as a whole rather than the individual questions and others made no reference to the specific questions while answering. Reference to the subscale questions could indicate a familiarity with the questionnaire gained during the interview or it may simply indicate an understanding of the question they were being asked and use of the information provided on paper for that subscale. Either way, the importance of having the questions on a sheet of paper for patients to refer back to during the interview was clear.

Often while thinking of the answer for one subscale a patient would be thinking about/describing other QOL concepts, particularly if a previous subscale discussed was deemed to be relevant to the next one. For some patients it was natural to place themselves on the subscales or answer the questions for themselves before thinking more generally about the group in the scenario, whereas others only thought of the questions abstractly rather than specifically for themselves. Earlier in the development of the interview we included a section where patients filled in the questionnaire themselves to gain familiarity with it rather than just hearing the description during the interview, it may be that this would have been useful for some but not for other patients.

Understanding of the questionnaire and scoring was demonstrated in a variety of ways during the interview. Some patients commented on or queried the time period the questionnaire covered and how you would fill it in if you had an odd off day on the day you were filling it in. Some patients were trying to understand how it works as everyone

starts off from a different point or at a different stage in treatment. Understanding of scoring was implied in different ways, mainly it was shown when the scores from the actual study were given and patients discussed what the graphs showed compared with their answers (sometimes interpreting the graphs before an explanation was given). Only one patient very clearly looked at the difference in scores from the study compared to the scoring system described for that subscale, referring to the sensitivity of the summary scores when one question moves one category. Confusion with the direction of scores arose in a few interviews but I generally could clarify this during the interview. Some patients commented on how confusing it was that for some scales an increase in score indicated improvement whereas for others an increase in score showed a deterioration in QOL.

Table 68 Familiarity with questionnaire and scoring

Major (number of interviews where extracted)	**Minor (number of interviews where extracted)**	**Quotes**
Familiarity (9)	Reference to subscale questions while answering (6)	Rather than tense and worried they'll be irritable and depressed (5 B p23)
	Linking subscales (4)	Sometimes it might be an emotional thing that you need to , that you know, that stops you from sleeping...or sometimes it could be a physical, that you've got discomfort (4 C p37) But with it being small (referring to pain subscale) they won't be, they won't have that much physical problems (9 A p27)
	Putting self onto scales (5)	

Major (number of interviews where extracted)	**Minor (number of interviews where extracted)**	**Quotes**
Understanding of questionnaire (7)	Querying or commenting on the timing of questionnaire or period covered (5)	It just depends exactly how you feel that day (2 p2) I'm sure if somebody had given me a questionnaire to fill out in the clinic when I was waiting for my diagnosis I don't think I would have filled it out properly I would have been able to concentrate enough to fill it out . I wouldn't have really thought about the questions, I might have ticked certain things but I wouldn't have really sat and thought about it (1 B p23)
	Awareness of multidimensional nature of QOL (4)	I presume if you can't carry things then it causes a problem and it causes you pain (1 A p8) You're not just dealing with the physical effects of the disease you're dealing with emotional effects of the disease the affect and sometimes the sort of guilt in a sense because you've got the pressure on your family around you and sort of affecting their lives in a sense so there's all these other things that come into it (4 p45)
	Showing an understanding of how QLQ-C30 may be useful (3)	

Major (number of interviews where extracted)	**Minor (number of interviews where extracted)**	**Quotes**
Understanding of scoring (10)	Confusion/clarity with direction of scoring indicating improvement/deterioration (3)	I've got it the wrong way round haven't I (1 A p8) It would be easy to get confused and mark them the other way (2 p26) I can see that that in a sense the reasoning behind them doing it now you explain it but it does seem to make it very complicated and to focus on which is which can sometimes maybe then it could lead to misinterpretation in a sense (4 p47)
	Reference to scoring system when thinking about answers (1)	I'd say there's no difference at all...because you're looking at sort of 20 points between each one and that's not even 10 (11 B p22) ..to shift from one box to another, because there's only four boxes so to actually move from one to another is quite a big difference, so I was thinking of that rather than on average everyone moving a little bit or a lot (11 p28)
	Understanding that your baseline score is important (4)	

Major (number of interviews where extracted)	Minor (number of interviews where extracted)	Quotes
	Suggesting specific score/category for group (8)	I thought that would be more like 60 (8 D p5) It would have started at not at all and it would go straight up to either quite a bit or very much (11 B p21)
Showing an understanding of the graphs (3)	Confirming answers vs graphs (2)	I thought I got quite right didn't I? (1 A p5)
	Interpreting before explanation given (3)	So theirs is better again isn't it? (6 C p15) So they could do most of it (1 A p8)

8.5.5.2 Summary: Can adequate familiarity with the QLQ-C30 and the way it produces quality of life scores be gained during an interview situation?

Patients demonstrated a familiarity with the questionnaire and scoring in a variety of ways. It was not clear that previous experience of filling in the questionnaire was particularly useful in gaining familiarity of the questionnaire for this exercise as patients generally did not recall whether they'd filled the QLQ-C30 in or something similar before. The interviewer needs to know the questionnaire, scoring system and scenarios really well as often the direction of scoring and of results (whether the difference represents an improvement/deterioration) needs clarifying during the interview due to function and symptom subscales being scored in the opposite direction. The link between the possible range of scores for a subscale and judging the meaning of a difference was not apparent in most of the interviews. It may be that by first asking the patients to judge expectation without seeing the scores they are less likely to consider the discrete nature of the scoring when judging the meaning of the scores from the study.

8.5.6 Secondary objective: To what extent can patients form their opinion using data from a group of patients rather than their own individual experience?

Two overall themes were found to relate to this secondary objective: 1) group or individual thinking, as planned in the original theoretic framework; 2) how patients approached the task, which emerged throughout the analysis of the transcripts. These were combined during the mapping and interpretation stage and are reported together here.

8.5.6.1 Group versus Individual

Text was extracted from the interviews where it was deemed to show evidence of patients thinking about a group of people or thinking of individual situations while considering their answers (Table 69).

It was relatively common for patients to refer to their own experience or to try and think how they would feel in the situation in the scenario if they had not experienced it. It was also clear that patients understood that they needed to think about a group situation, with all patients mentioning the variation of people within a group; 10 patients clearly referring to a group situation in their answers. Six patients referred to how hard it was to think about the average from a group rather than an individual. Some patients discussed their experience of others or thought more in the abstract about the group while thinking of their answers.

8.5.6.2 How patients approached the task

How individual patients approached the task emerged as a new theme while reviewing the transcripts. It is summarised here under two minor theme headings: 1) using their own experience then generalising to a group; 2) no reference to their own experience. It shows that although most patients had a tendency to mention or think about their own experience they were also putting this into the context of the group in the scenario.

Table 69 Evidence of group versus individual thinking

Major (number of interviews where extracted)	Minor (number of interviews where extracted)	Quotes
Group (11)	Awareness of variation in group (11)	Some might not be able to take...some of us can take pain quite well others speak out first (1 A p6) There's such individual variations about it, you know, some people it would affect them in a negative way in a sense and other people it could you know it could be very positive for them (4 C p22) And of course you get such variation from people don't you (7 B p10) We've all got different anxieties that keeps us from sleeping (9 C p10)
	Clear group thinking (10)	I was thinking back to myself because obviously you just think how did I go through it and you just think well on a scale of average I must be just average because I'm just someone else (11 p27) I would feel as a general group... (4 A p11) so tempted to just answer for myself, I mean I can't you don't want me to do that do you? You want me to think about the group? (2 C p6)

Major (number of interviews where extracted)	Minor (number of interviews where extracted)	Quotes
	Reference to others in the abstract (7)	I had pain but it didn't stop me from doing anything...And I'm sure other people have had pain and it has stopped them (5 A p6) Some people might feel quite poorly (6 D p9)
	Reference to experience of others (4)	I was absolutely fantastic ... I know friends that I made while I was having chemotherapy that were very ill and sick and terrible (5 A p5) Some people carried on working, I just felt so poorly (2 C p15)
	Thinking of majority (4)	I mean if you say medium that covers most, so medium I would say, but you will have the isolated large (10 A p16) I think it would kind of vary between a little and quite a bit ...and you'd get the odd one or two that says very much (11 C p5)
	Finding average difficult (6)	It's so difficult when you've got to do it, it's getting my head round just this average group... because everybody's an individual (2 C p14) it's very difficult to know because individually there will be people who respond differently (4 C p21) It's hard to think in groups isn't it?... When you know individuals and things. (10 B p11)

Major (number of interviews where extracted)	**Minor (number of interviews where extracted)**	**Quotes**
Individual (9)	Reference to self in the abstract (4)	if it me.. not been down that road my feelings would be ...(1 A p4) I'm thinking, you know, if it's me at this stage (8 p17)
	Reference to own experience (7)	it still kind of put my own perception on it rather than focussing on knowing what a 50 year old woman's kind of going through. (11 p27) I could still do everything like that (1 B p13) I think to be honest it's got to be how I experienced it and I certainly was very limited with work and my daily activities (2 C p7)
	Answering subscale for self (3)	I had no trouble sleeping, not on chemo (9 C p10)

Major (number of interviews where extracted)	Minor (number of interviews where extracted)	Quotes
Approach to task (11)	Using own experience then generalising (10)	I just go on gut feeling and me own scenarios...you know what I've been through and sort of try to put my experiences in some sort of middle ground, because we're all different and how I would I would look at it and expect them to feel. (1 p27) I know it obviously affects different people differently, some people you know, I think have been more affected than me and other people you know, much less really. I think that depends on what's the matter. (6 p23) I was thinking back to myself because obviously you just think how did I go through it and you just think well on a scale of average I must be just average because I'm just someone else (11 p27)
	No reference to own experience (2)	So I think probably for this first lot no because I mean slight arm pain, I think, I don't think it would impinge on doing any of these like walking and things (4 A p14)

8.5.6.3 Summary: To what extent can patients form their opinion using data from a group of patients rather than their own individual experience?

All patients seemed able to grasp the concept of thinking about a group situation rather than an individual one and were aware of the need to do this during the interview. However, it was common for patients to mention it was difficult to think about an average. Some clearly went on their own experience or preference to form opinions

and a minority seemed to form an opinion just on the information in the scenario, approaching the task in a more generic and abstract way. Even if patients did use their own experience to form an opinion they seemed to try to think about their experience versus others in order to extrapolate to an answer for a group rather than an individual.

8.5.7 How do patients' opinions compare with clinicians opinions when using the same published data?

Three experts reviewed each of the scenarios. Scenarios A and D were also reviewed by five patients and scenarios B and C by six patients.

8.5.7.1 Quantitative results

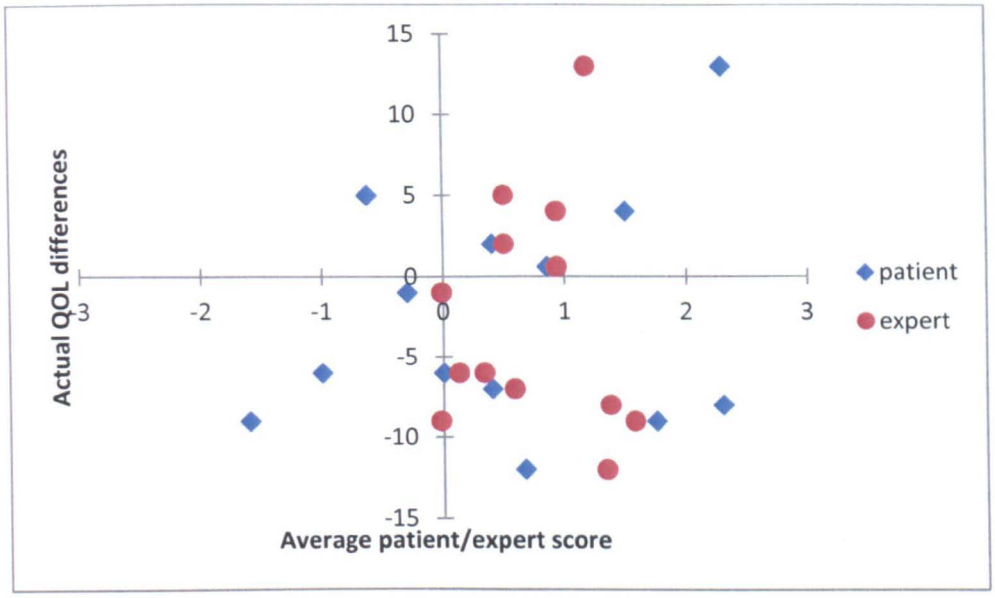
The maximum difference between patients and experts was 1.1, which equates roughly to a difference of one size class (Figure 57). There are two instances where patients and experts judged the differences in opposite directions. Whilst there was sometimes wide variation in the expert scores there were no disagreements between experts in the direction of expected change for the contrasts used in the scenarios. However, there were two of the contrasts on scenario D that led to patients having expectations in opposite directions from each other.

Where patients and experts agreed on the direction of the change (11 contrasts), the patients' opinion was larger in magnitude for six of the contrasts and smaller for five of the contrasts.

Figure 57 Expert versus patient opinion

Figure 58 shows the average patient and expert scores for each contrast compared with the actual scores from the papers. The correlation between patient scores and the actual scores was 0.25. This is similar to the correlation seen in the overall study between the experts and the actual scores for longitudinal contrasts (0.28). However, the correlation between the average expert scores and the actual scores for the subset of contrasts used in this pilot study was much lower (-0.01) than for the patients. Note that I deliberately chose some contrasts where the experts had disagreed with each other in order to test these contrasts on patients however this is still similar to the correlation seen in the overall study for the cross-sectional contrasts (-0.03).

Figure 58 Correlation of average patient and average expert scores versus actual scores



8.5.7.2 Qualitative results

Generally patients made few comments on the differences between their judgements and those of the experts, even if they were surprised by the differences. Those who did comment talked about why there may be differences or agreement between them and also commented on the ability of clinicians to judge on QOL differences. A summary of the comments are displayed in Table 70 under three major sub-themes grouping the comments into those about the ability of clinicians to judge patients' QOL and comments relating to why there may be agreement or disagreement between the patients and experts.

Table 70 Patients versus clinicians: patients' comments

Major (number of interviews where extracted)	Minor (number of interviews where extracted)	Quotes
Ability of clinicians to judge (5)	Able (4)	If they're seeing them in clinics then they would, you would definitely know because even if the patient said oh no I'm fine, you'd see how they were sitting and how they came in and that sort of thing (11 B p23) I also think doctors are very very good because they can know who people, what they need really, and I think they can soon weigh you up (9 C p10) I mean the doctors should know because they keep the notes and they see everybody don't they so between them they should be able to come up with the answers (5 A p14)
	Not able (1)	Well they're probably looking at it from the point of view that they, the doctors and nurses, have actually done something about this problem and therefore in their eyes have, not solved it entirely of course, but they've gone, they've cut out the immediate one, you know they've got over the first hurdle... But that's not the story really, your first hurdle is your first hurdle and then people then go on to do other things (10 B p13)
Reasoning for differences (4)	Can use their experience (1)	Or is it that this pain is manageable and do the doctors know through their experience and the amount of people they've seen with it realise that yeah they might be in a little bit more pain but not much (1 A p6)

Major (number of interviews where extracted)	Minor (number of interviews where extracted)	Quotes
	Experts may assume positive outcome (1)	They're probably looking at it from the point of view that they, the doctors and nurses, have actually done something about this problem (10 B p13)
	Place different importance on QOL changes (1)	
	Can see reasons why differences could be either way round (1)	...think it just depends on how, if you were feeling really well before or whether you were having some effects really (6 D p9)
Reasoning for agreement (1)		But I suppose I'm in some way imagining what the doctors and nurses are telling me you know, and the information that you're given as well, I haven't experienced it yet (6 C p20) ...a lot of it I'm going off what the doctors and nurses have told me as to how I imagine that I'll recover from the chemotherapy really (6 p22)

8.5.7.3 Summary: How do patients' opinions compare with clinicians' opinions when using the same published data?

The maximum difference between the average patient and expert scores was only one size class. Patients were just as likely to judge differences to be larger or smaller when compared with the experts opinions. The patient opinions correlated more highly with the actual scores in this subset of contrasts than the expert opinions.

8.5.8 Does the proposed interview need developing further prior to a larger study? Is the information presented in a way patients can understand?

8.5.8.1 Scenario development during the course of the interviews

Interviews were reviewed after every three interviews to see if patients were asking for information that was not available in the scenario summary. During the course of the interviews all the scenario summary sheets were amended slightly as patients raised queries about the study or patients in the study. The majority of changes were made to include extra treatment information and this makes sense given that it was evident that the majority of patients (10/11) used treatment information to inform their decisions. Other information that was discussed as useful to have in the scenario was:- type and severity of pain (one patient), other QOL dimensions (one patient referring to the fact that sleep depends on emotional, pain and symptoms).

8.5.8.2 Patient definitions of small, medium and large differences

During the interviews patients were asked to define the size of the difference they expected, and the wording suggested by the interviewer was 'small', 'medium' or 'large'. In order to understand how patients approach the task and to see if the wording for size classes should be further developed, patients were asked at the end of the interview whether they had specific definitions of small, medium and large differences in mind when deciding upon a size class for each contrast. Data was also extracted and is summarised here to show more generally the wording patients were using themselves while working out the differences. Generally, when asked to define what they were thinking when deciding between small, medium and large, patients only gave a definition of what they thought would make a difference large. No common definition emerged. While working through the scenarios patients used a wide variety of terms to think of the size of the difference but most commonly used the categories from the QLQ-C30, i.e. a little, quite a bit etc. One patient commented that the wording used in the interview was irrelevant really. It was also common for patients to say they just went on a 'gut instinct' for each rather than having a specific definition in mind.

Table 71 Description of small, medium and large differences

Major (number of interviews where extracted)	Minor (number of interviews where extracted)	Quotes
Specific definition of small (1)	Treatment working; no pain and able to function (1)	Like I was last year when the tablets were working and I felt wonderful and I didn't have any pain, you know and I could do all those things like lift bags and that (1 p28)
Specific definition of medium (2)	Sizeable (1)	A sizeable difference (11 C p8)
	Ups and downs (1)	Medium is like I always think I am on a you know mostly bobbing about but then having a few dips and then go up (1 p28)
Specific definition of large (5)	Extremes (1)	...in extreme pain, not being able to do anything, being totally depressed and not being able to do anything (1 p28)
	Own experience vs others (1)	I think a large difference is when, I mean I was fine and when other people were being sick and feeling really really tired, I think that's a big difference (5 p27)
	A little vs A lot (1)	Trying to get in to sort of the mind of someone to think what would make them change from answering 'a little' to 'a lot' or to 'completely'. (11 p27)
	A marked difference (1)	...so I just thought there just be a marked difference between me and the other person (8 p19)
	Significant difference on your life (1)	It could make a large difference in a sense if you've got lots of physical things that are affected by it kind of thing and or if you find it emotionally very difficult to deal with then again it can have quite a significant difference on your life so it was trying to sort of think of that (4 p44)

Major (number of interviews where extracted)	Minor (number of interviews where extracted)	Quotes
Using questionnaire responses (8)	A little, quite a bit, very much (8)	I think it would kind of vary between a little and quite a bit (11 C p5)
Other wording	Percentages(3)	Let's say 15 out of 20 do well here compared to 3 or 4 out of 20 here (8 C p12) I would think medium to...55%... yeah (3 C p15)
	Marked (2)	I think there'd be a marked difference between the two groups because it would be obvious (8 C p11)
	Slight (2)	There's a slight difference but I wouldn't say a lot ...I wouldn't say a marked difference (1 A p9)
	Average (1)	It's hard to put maybe average I suppose I don't know (4 A p9)
	Substantial (1)	I think it might be quite substantial (4 A p9)
Wording irrelevant (1)		I think you automatically picture in your mind anyway it's the same thing however its worded (3 p25)
Not specific (7)	Gut instinct (7)	How I felt for each really (7 p22) just gut instinct really of the sort of medium, just somewhere in the middle really isn't it, and that either side of that would be the small or large bit (6 p22)

8.5.8.3 Impact of showing patients the actual scores

In the majority of cases (63% of contrasts judged in the interviews) the patient did not change their judgment after seeing the scores. On average the patients reduced the size of their scores by a mean of 0.1 points (median of zero) once they saw the actual mean scores. There were only two instances where patients decided to change

the direction of their original judgement and only three instances (out of 19 changes) where they increased the size of their score after seeing the results from the study.

8.5.8.4 Is information presented in a way patients can understand?

During the interviews there were a number of occasions where clarification was required from the interviewer either on the details of the scenario or on the question being asked. Some of the scenarios had a lot of information in them and forgetting one fairly small detail may have had a big impact on the results. For example, one patient was discussing scenario C and referring to how the support group would help after diagnosis but the setting was in metastatic breast cancer on average 5 years after diagnosis. Generally, queries or misunderstandings were easy to resolve during the interview by reminding the patient of some of the details. Sometimes it was hard to steer the patient into judging the difference required between groups or over time either because they were talking about the individual groups or were trying to answer the subscale questions for themselves.

Patients seemed to understand the graphs and some commented that they found them useful to help visualise the group. Sometimes there was initial confusion with the direction of the change shown by the graphs but this could be easily clarified in the interview. A number of patients interpreted the graphs before an explanation was given, a lot of them commenting on how they thought the graphs compared to their answers, e.g. higher or lower than expected. However, I was concerned that sometimes there was an over-reliance on the graphs, which were designed to show what an average score might look like rather than the actual results from individuals in the study. In particular, patients sometimes highlighted individual patients as standing out on the graphs and swaying their decision.

Table 72 Presentation and understanding

Major (number of interviews where extracted)	Minor (number of interviews where extracted)	Quotes
Confusion with details of scenario (3)	Forgetting timelines of patients from diagnosis (1)	Oh right, well that puts a different reflection on it again. (9 C p11)
	Patient referring to pain but scenario is regarding symptoms (1)	I can't understand why one lot of people are saying that they haven't had any pain (5 A p6)

Major (number of interviews where extracted)	**Minor (number of interviews where extracted)**	**Quotes**
	Patient referring to normal care group as having no support (1)	I think to some extent they should be limited because they are alone they don't have any support whatsoever (8 C p11)
Confusion with required task (4)	Needing clarification of difference rather than by group (3)	Is this for group one then first we're doing? (1 A p7)
	Trying to complete task with reference to self rather than scenario (3)	Problems in the arm, oh this is in the last month, well I, ...no definitely not stiff, no, no pain, no, I've had no infections. (5 A p6)
	Not understanding purpose of scoring explanation (1)	Can I just ask why we need to know the scores? (5 A p9)
Graphs (9)	Own interpretation (8)	I'm surprised, I think that is quite good, higher than I thought it would be (9 C p9) It's getting back to how they were prior to chemotherapy (6 D p9)
	Confusion with direction of difference (2)	I'm getting it the wrong way round to the, oh of course, this is the best bunch so in effect...(4 C p24)
	Influenced by individuals/extremes on graph (5)	It's obviously noticeable for one person (6 C p19) when you look at the graph you've got a little person, you've got person at the very beginning of the graph and somebody at the end (2 p25)
	Useful (3)	The graphs were brilliant because it shows you the shift (11 p29)

Major (number of interviews where extracted)	Minor (number of interviews where extracted)	Quotes
Understanding of scenario (9)	Linking subscales and using previous subscale answer (1)	So they've only got a little bit of pain, so they can do most things (1 A p9)
	Awareness of timescales in scenario and impact (5)	It can often take much more than 8 weeks really but of course we are speaking about the 8 week period (10 B p12) It's not as if they've just had surgery or something anyway is it where your ability's not that good anyway and then over a short space of time it gets a lot better(11 C p13)
	Showing in-depth understanding of groups in the scenario (5)	Is it because there's a lot of being in the support group that their activities may be limited? (8 C p10) I was going to ask as well you know this relaxation exercise, did group two know about this exercise, did they have it in a like a leaflet or anything or was it just given to the support group? (11 C p4)
	Showing in-depth thought around subscale (3)	If you've got a stiffness in an arm and depending on which...whether it's your sort the arm you used most of all you know if you're left handed or right handed that obviously could affect... (4 A p14)

8.5.8.5 Summary: Development for future interviews

The scenario summary sheets were improved over the course of the interviews, adding extra information where it was deemed to be required. This was mainly the addition of treatment details or further clarification of the groups in the scenario.

Scenarios C and D may have been lacking in some of the information patients required as this information was not available from the paper.

The wording used in the interview to describe differences (i.e. small, medium or large) is probably sufficient, although sometimes patients judged in between these categories. The other wording that could be considered is based on the QLQ-C30 question responses of 'not at all', 'a little', 'quite a bit' and 'very much'.

Seeing the actual scores from the study had a small impact on patients' opinions. Where patients did change their mind the tendency was to reduce the size of the QOL difference compared to their original expectation.

Patients seemed to understand the information presented in the interview although there is a lot of information to remember and reminders of the scenario details were sometimes required from the interviewer. The concept of looking at the difference between groups at a certain point in time rather than individual groups over time was sometimes hard for patients to understand. Although the graphs were highlighted as useful and patients seemed to be using them to compare with their original judgements, care needs to be taken in how the distributions of individual stick figures on the graphs are drawn as there was some evidence that sometimes stick figures at the extremes unduly influences some patients' judgements.

8.5.9 Which types of scenarios should be developed for a further study?

8.5.9.1 Which scenarios were easier?

As described previously, each patient considered two scenarios during the interview. Patients were asked if they found one scenario easier to think about than the other. Only two people thought they were of equal difficulty, both of these had scenarios C and D. Most patients (9/11) found one scenario was easier than the other. The scenario most consistently appearing easier was scenario B which five patients identified as easier. Generally people seemed able to relate to this as it was regarding clinics and waiting for diagnosis. The scenario most consistently appearing harder was C, which was never chosen as the easiest scenario. This scenario was looking at patients attending a support group compared with normal care. Although scenario C had further information added as the interviews progressed (average age, time from diagnosis and treatment information) it was still identified as harder, either because patients found it hard to relate to without having experienced a support group of that

nature or because they felt that whether a patient benefited from it or not was a very individual thing.

Table 73 Easier/harder scenarios

Interview	Easier scenario	Harder scenario	Reasons (summarised from interview text)
1	B	A	Can ask any cancer patient scenario B as all been in a clinic. Scenario A was harder because she didn't have arm problems. (p26,28)
3	B	D	Scenario B, the stress factor, easier to imagine how people would respond. (p25)
4	A	C	Scenario A easier as looking at physical things, even though she hasn't experienced it exactly as in scenario. Scenario C harder as more information needed, it was very individual as to whether patients would benefit or not. (p22,41)
5	B	A	Couldn't get head around A (p26)
6	D	C	D easier because relevant to the treatment she's had, also could imagine herself and how she would feel if in the older age group. C didn't have enough information – needed treatment details (p21,22)
7	B	D	D needs more information, age alone is not enough, there are a wide variety of people within an age group eg in terms of physical fitness which effects how well you do. (p14-16,21)
9	A	C	Scenario C is a very individual thing. Scenario A, pain and physical are logical and can relate to this more. (p30,31)
10	A	B	Scenario A easier because she was a nurse (now retired) and because of her experience of talking to people. (p18)
11	B	C	Scenario B easier as can relate to how they'd be feeling in the clinic situation. Scenario C harder as not been in a support group and also not at that stage of disease. (p26)

Patients raised a few difficulties relating to specific scenarios (Table 74). Difficulty in judging QOL differences arose where there was a broad group of patients in a study or where the effect of the intervention or situation could vary greatly between patients. Certain scenarios were harder without experience, in particular scenario A was highlighted as hard to judge without experiencing arm problems yourself.

Table 74 Difficult scenarios

Major (number of interviews where extracted)	**Minor (number of interviews where extracted)**	**In relation to which scenarios (number of interviews)**	**Quotes**
Identifying difficult scenarios	Due to variation in group or differing effects on individuals (4)	A (1) B (1) C (2) D (1)	It's a bit difficult because it really does depend on the person, it could be quite large in some cases (10 B p6)
	Due to lack of experience (4)	A (3) D (1)	...having never experienced that ... it's hard to know in affect just how it would feel, how severe it might be (4 A p15) But not having been there yet, it's quite hard you know, it's just an assumption really (6 D p8)
	Due to their experience not being the norm (1)	D (1)	I mean I have some friends and they really don't get involved with anything ... I mean I was playing badminton at 77 you know and walking with them you know coast to coast, all sorts you know long walks...So and this is all since I retired which is 60, you know so, you know it's difficult for me to judge (7 D p15)

8.5.9.2 Are any of the subscales easier for patients to judge?

Six patients felt there was no difference between the subscales in terms of the difficulty in thinking about QOL changes on that subscale. Four patients felt there were differences in difficulty and one patient was not asked to comment on the difficulty of the subscales (this was an omission on my part). All four patients highlighted the physical functioning subscale as easier, one also mentioned fatigue and one also mentioned pain as easier. Role functioning and sleep subscales were described as harder because they are very individual by one patient and similarly emotional functioning was described as an individual thing by another patient.

8.5.9.3 Summary: scenarios for future studies

These findings would suggest that studies with which the majority of patients can relate to (e.g. scenario B looking at different types of clinics for diagnosis) are the most useful to use for this purpose. These are easier for patients to judge the expected QOL changes. Scenarios relating to specific symptoms or problems may be best aimed at people with similar experience to that in the scenario rather than patients who have not experienced those or anything similar. Scenarios need to use studies where the groups are well-defined and not too varied in terms of disease or patient characteristics. Studies of psychosocial interventions (e.g. scenario C support group therapy) may not be useful as the effect of the intervention can vary so greatly depending on the individual. No subscales were highlighted as too difficult to judge although only seven out of the possible 15 have been used in these interviews.

8.6 Validity

Four transcripts were independently reviewed and coded by the rest of the study team. Only minor discrepancies in coding were noted on one of the transcripts and these were amended. These would not have changed the interpretation of the results as they only changed coding at the level of the minor themes.

8.7 Patient opinions - Conclusions and discussion

The study showed that patients were able to do the required task and there was some indication that their opinions could be more closely related to the actual scores from the original papers than the expert scores were. However, the agreement

between patients was much poorer than the agreement between experts, for this subset of contrasts and also compared with the full set of contrasts. For future studies it may be necessary to consider a consensus process if patient opinion is to be used to develop interpretation guidelines.

Patients recognised the need for a mix of patients to review the contrasts due to wide variety of experiences and opinions they may have. It was also apparent that patients rely on their own experience of the disease and treatment in making their decisions. Patients' values and personality may also influence how they judge the QOL differences. For these reasons a patient panel may have to be larger than the expert panel in order to capture the wider range of opinions.

For future studies, careful consideration needs to be put into the types of studies that can successfully be used as scenarios. Studies need to have a sufficient amount of detail in terms of patient and disease characteristics and have a study population that is not too diverse. The format of these interviews worked well but a decision would need to be made up front as to whether patients were to carry out the task with or without the knowledge of the QOL scores. Although we tried to address which was better in this study, I found that once patients had made their decision on a blinded basis they were unlikely to change their scores after seeing the actual QOL changes. We cannot tell from this study if showing the patients the QOL scores first would have resulted in better agreement.

Although there were issues with the patient reviews, most notably the lack of agreement between patients, this study succeeded in presenting the information in a way patients could understand and obtained patient opinion on the size of QOL differences. The description of the QOL questionnaire and mechanism for obtaining a score seemed to be well understood and enabled patients to make decisions on QOL differences from the study scenarios. Patient opinions should be considered alongside expert opinions for future studies as it was apparent that they offered a different perspective and insight as to the likely size of difference in QOL scores. I would recommend an initial training program for the patients, similar to that carried out for the expert panel. However, rather than patients then working remotely, better quality reviews are likely if the patients and experts work together. This could be as part of a focus group in order to reach a consensus or reject contrasts where opinions are too varied.

9 Overall Conclusions and Discussion

Sections of this chapter have been based on the Discussion in our Journal of Clinical Oncology paper (full reference below). The text has been re-written here with further details and new discussion points, particularly since the original article only reported the cross-sectional guidelines.

Kim Cocks, Madeleine T. King, Galina Velikova, Marrissa Martyn-St-James, Peter M. Fayers and Julia M. Brown. "Evidence-Based Guidelines for the Determination of Sample Size and Interpretation of the European Organisation for the Research and Treatment of Cancer Quality of Life Questionnaire Core 30" *Journal of Clinical Oncology (Jan 2011), Number 29, Issue 1, p89-96.*

9.1 Introduction

Quality of life questionnaires are now widely used in health services research but the question around how to interpret QOL scores has been the subject of much debate. Although a number of different methods have emerged to try to find the minimally important difference between QOL scores there are a lack of guidelines that are specific to the questionnaire being used.

The EORTC QLQ-C30 is one of the two most widely used questionnaires to measure QOL in cancer patients and the focus of this research. When I conducted a systematic review of the RCTs reporting QLQ-C30 scores I found that, although the standard of reporting was high, the clinical meaning of QOL changes was rarely addressed. If clinical significance was addressed, there was no clear standard used for the interpretation. This research therefore aimed to produce interpretation guidelines specifically for the QLQ-C30.

Given that the questionnaire had been in use for a number of years we used a novel approach in order to develop guidelines using the rich evidence-base available from published studies reporting QLQ-C30 scores.

9.2 Summary of study and results

The study consisted of a literature search to find all sources of mean scores from the QLQ-C30 followed by a meta-analysis to estimate the size of trivial, small, medium

and large differences. A panel of experts were used to group the published mean scores into the size classes for the analysis. I also conducted interviews with patients to see if we could develop the methodology further to incorporate patient opinion into the guidelines in the future.

The study succeeded in producing evidence-based interpretation guidelines, based on high quality studies, for trivial to large QLQ-C30 QOL differences (Table 75 and Table 76). The novel methodology for deriving the guidelines was found to be a robust and valid approach following in-depth analysis and qualitative work carried out with both experts and patients.

Table 75 Guidelines for size of cross-sectional differences

Threshold between small and medium estimates	Sub-scale	Mean difference				Effect size			
		Triv	Small	Medium	Large	Triv	Small	Medium	Large
<10 points	DI	0 - 3	3- 7	>7	-	0-0.1	0.1-0.4	>0.4	-
	NV	0 - 3	3- 8	8 - 15	>15	0-0.2	0.2-0.5	0.5-0.8	>0.8
	CF	0 - 3	3 - 9	9 - 14	>14	0-0.2	0.2-0.4	0.4-0.7	>0.7
	DY	0 - 4	4 - 9	9 - 15	>15	0-0.1	0.1-0.3	0.3-0.6	>0.6
10-15 points	FI	0 - 3	3 - 10	>10	-	0-0.1	0.1-0.4	>0.4	-
	QL	0 - 4	4 - 10	10 - 15	>15	0-0.2	0.2-0.4	0.4-0.6	>0.6
	SF	0 - 5	5 - 11	11 - 15	>15	0-0.2	0.2-0.4	0.4-0.6	>0.6
	SL	0 - 4	4 - 13	13 - 24	>24	0-0.1	0.1-0.5	0.5-1	>1
	FA	0 - 5	5 - 13	13 - 19	>19	0-0.2	0.2-0.5	0.5-0.8	>0.8
	CO	0 - 5	5 - 13	13 - 19	>19	0-0.2	0.2-0.5	0.5-0.8	>0.8
	PA	0 - 6	6 - 13	13 - 19	>19	0-0.2	0.2-0.5	0.5-0.8	>0.8
	PF	0 - 5	5 - 14	14 - 22	>22	0-0.2	0.2-0.6	0.6-1	>1
	AP	0 - 5	5 - 14	14 - 23	>23	0-0.2	0.2-0.6	0.6-1	>1
>15 points	RF	0 - 6	6 - 19	19 - 29	>29	0-0.2	0.2-0.7	0.7-1.1	>1.1

Table 76 Guidelines for size of longitudinal differences

Threshold between small and medium improvements (deteriorations)	Sub-scale	Deteriorations			No difference	Improvements		
		Large	Medium	Small	Trivial	Small	Medium	Large
<10 (>-14)	FI	NE	<-10	-10 to -2	-2 to 3	>3	NE	NE
	CF	NE	<-7	-7 to -1	-1 to 3	3 to 7	>7	NE
	PF	<-17	-17 to -10	-10 to -5	-5 to 2	2 to 7	>7	NE
	QL	<-16	-16 to -10	-10 to -5	-5 to 5	5 to 8	>8	NE
	SF	NE	<-11	-11 to -6	-6 to 3	3 to 8	>8	NE
	EF	NE	<-12	-12 to -3	-3 to 6	6 to 9	>9	NE
	NV	<-16	-16 to -11	-11 to -5	-5 to 3	3 to 9	>9	NE
	DY	NE	<-11	-11 to -5	-5 to 2	2 to 9	>9	NE
	FA	<-15	-15 to -10	-10 to -5	-5 to 4	4 to 9	>9	NE
	SL	<-17	-17 to -9	-9 to -2	-2 to 5	5 to 9	>9	NE
	PA	<-20	-20 to -11	-11 to -3	-3 to 5	5 to 9	9 to 14	>14
≥10 (≤-14)	CO	NE	<-15	-15 to -5	-5 to 4	4 to 10	>10	NE
	DI	NE	<-15	-15 to -5	-5 to 3	3 to 11	>11	NE
	RF	<-22	-22 to -14	-14 to -7	-7 to 6	6 to 12	>12	NE
	AP	<-26	-26 to -14	-14 to -2	-2 to 7	7 to 13	>13	NE

9.3 Conclusions and discussion

The guidelines utilised almost 3500 contrasts from published scores and incorporated reviews from 34 cancer and quality of life experts. These new estimates highlight that previous guidelines may be too simplistic in that they do not distinguish between subscales and are applied to both longitudinal and cross-sectional differences. We add to the current literature by providing separate guidelines for interpreting differences arising from between group comparisons and those measured over time. The guidelines are also specific for each QLQ-C30 subscale. Researchers

can now more accurately calculate sample size according to the subscale of primary interest.

Our study also indicated that deteriorations over time are generally larger than the corresponding estimate of an improvement for the same subscale, therefore a global rule for a clinically meaningful change is not appropriate. Cella *et al*(86) also observed that patients indicating a change for the better using the FACT-G questionnaire were more responsive to small changes whereas those experiencing a worsening required a somewhat larger degree of change. Our longitudinal guidelines therefore allow specific interpretation according to whether scores improve or deteriorate over time.

Both the longitudinal and cross-sectional guidelines are more widely applicable than those currently available as they are based on more than 200 papers from a wide range of cancers and clinical situations. Previous work from Osoba *et al*(49) was based on around 350 patients from two studies of breast and lung cancer and King *et al*(19) used 14 cancer studies.

9.3.1 Recommendations for practice

We suggest the threshold between trivial/small in our guidelines would be the smallest estimate on which to base a sample size. Depending on the individual study and the type of interventions larger differences may be of interest and the range of small or medium estimates could be used. Our study showed that it was rare for the experts to expect large differences between groups even when comparing very distinct groups of patients. Of particular interest when designing clinical trials is the fact that if we retrospectively apply our cross-sectional guidelines to the contrasts comparing treatments from RCTs, large effects are observed in only 14 (2%) of the contrasts. Also, large differences over time were almost non-existent. This lack of large differences has also been observed in other studies of patients over time(49;86). Therefore, researchers designing a clinical trial should consider at the outset if large effects can reasonably be expected. Observed changes in the QLQ-C30 subscales over time are likely to be small in most clinical situations as seen in our observed differences. This could be due to response shift in patients filling in the questionnaire over time(97). Care should be taken in planning studies where the primary aim is to observe a change over time in QOL measured by the QLQ-C30.

Compared to King(19) our results for small effects are very similar for physical, role, cognitive, nausea and pain. For the cross-sectional contrasts our estimates of medium effects lie in the same range as suggested by Osoba *et al* (10 to 20 points) for

global QOL (10-15 points), social (11-15 points), pain, constipation and fatigue (13-19 points) subscales. For certain subscales (e.g. NV and DI) these commonly used guidelines will under-estimate sample sizes and may also miss important differences in QOL when used for interpretation. At the other end of the spectrum, our guidelines indicate that sample size calculations using the RF subscale in particular will be over-powered if using the 10 point threshold. Our guidelines indicate that the importance of changes on the RF subscale may currently be over-estimated, for example a difference of 20 points would be interpreted as moderate/large using Osoba's guidelines but is small/medium using our guidelines.

For the longitudinal contrasts our thresholds between small and medium improvements were mainly smaller than Osoba's 10 point threshold (7 to 9 points for 10 of the subscales). The medium deteriorations on the other hand had thresholds of 10 points or more for 13 of the subscales.

Our study and King's considered group differences using published data whereas Osoba used individual patients' ratings of change over time to produce guidelines. Despite these differences, there was substantial overlap in the resulting guidelines from the three studies.

There were also similarities in the meta-analysis estimates between the EBES project for the FACT-G and our meta-analysis estimates. Although the items are different in the two questionnaires there is some overlap between the subscales and domains. The physical well-being domain from EBES had effect size estimates of 0.42 and 0.87 for small and medium cross-sectional effects respectively compared to our physical functioning subscale with estimates of 0.41 and 0.84 respectively. The functional well-being effect size estimates were a little smaller than our role functioning estimates (0.37 and 0.71 for small and medium in EBES versus our estimates of 0.46 and 0.84). The emotional well-being estimates were 0.32 for small effects and 0.40 for medium effects in EBES. Our small effect size estimate for emotional functioning was similar, 0.35, but the medium effect was lower (0.28). The social/family well-being estimates from the FACT-G were much lower than our estimates for the social functioning subscale (0.14 and 0.23 versus 0.38 and 0.46 respectively for small and medium estimates).

9.3.2 Limitations and recommendations for further research

We used experts to estimate impact on patients' QOL. Experts have previously been shown to under-estimate symptom severity(98). However, the same study also

showed that in the context of an RCT, patients and doctors had similar conclusions with respect to between treatment differences with respect to physical symptoms. We believe the use of experts was justified here as we were seeking to quantify the size of differences on groups of patients from clinical studies rather than estimate an individual's QOL. We chose experts familiar with the QLQ-C30 so they could use their knowledge of the specific questions as well as clinical experience.

It is important to note that some of the issues found in our study were not unique to our methodology and similar anomalies were also found by Osoba *et al*(49) who used patient ratings. They found the emotional functioning subscale had an unusually large difference estimate for the no change category and actually had an increase in score for 'a little worse'. We also found issues with the emotional functioning subscale, where medium differences had a lower estimate than the small estimate. In Osoba's study physical functioning had very little change in the 'a little better' category compared to the other subscales. The global QL scale had moderately worse and very much worse scores of only 3.8 and 5.6 respectively. Similarly in our study PF and QL have amongst the lowest thresholds between small and medium estimates in both directions.

I conducted the pilot study to research feasibility of incorporating patient opinions on published data in a similar way. Although patients could gain some understanding of the instrument and scoring, they found it hard to judge differences for groups of patients. They generally relied on their own experience and that of a few people around them. However, there was also some indication that their opinions could be more closely related to the actual scores from the original papers.

The patients' opinions were quite varied and the agreement between them was poorer than for the experts. For future studies I would recommend having a larger panel comprising of both patients and experts but to include a formal procedure for obtaining consensus on opinion. There may need to be more patients than experts on the panel in order to obtain the wide range of experience required. Since experts and patients average opinions were only a maximum of one size class apart it seems reasonable to assume that a consensus process could work well in order to combine the expert and patient opinion. In conclusion, I believe that although the inclusion of patients in the process is desirable, our use of an expert panel here resulted in acceptable size classes to group the contrasts for the analysis.

A separate issue with the expert review is whether the experts could truly be blinded to the QOL results. This is particularly pertinent in cancer sites with only a few

papers reporting QOL or for the larger more well-known studies. It is likely that an expert in the field would already be familiar with the study results. We relied on the experts own judgement and asked them to return papers if they felt they already were familiar with the QOL results. This may be an area where involving patients in the panel could improve the methodology as they are very unlikely to be familiar with the medical literature.

In the similar FACT-G project(1) contrasts with agreement between the experts were highlighted as important for validity of results. However, it was unclear which study or contrast characteristics led to good agreement. Therefore for this study we did not set stringent criteria for inclusion but later applied criterion to exclude what we felt were the poorer quality contrasts. Post-hoc exclusion is a weakness in our study and future studies would benefit from excluding these contrasts up front. A large proportion of contrasts undergoing the review process were subsequently excluded from the analysis, either due to lack of agreement between the experts or disparity in the direction of the expert review compared with the actual results. Further studies could improve on efficiency by using our analyses to exclude contrasts at the start of the study which we subsequently found led to poor expert agreement or correlation with actual scores (such as heterogeneous patient groups or groups of patients with a mix of disease stages for example).

More contrasts undergoing the expert review could also have been included if we had used a full consensus process between the experts, with experts communicating with each other to reach a consensus where possible. We contacted the experts individually where there was a discrepancy in their ratings and asked them to check their reviews for errors. We did not attempt to bring experts together in order to reach consensus as this was deemed unfeasible with the large number of reviewers on the expert panel spread internationally. King *et al(1)* had previously used a similar system with greater success but only had three expert reviewers on the panel who reviewed all of the papers. I would recommend a more thorough system for obtaining consensus for future studies using this methodology in order to include more of the identified published data.

The poor concordance between experts may also have been due to the complexity of the rating scale. The experts had to understand the direction we intended on the score sheet when referring to negative and positive scores. This was particularly complex for the cross-sectional contrasts where the direction depended on which way round we considered the groups and this may have led to some ratings being placed in

the wrong direction. Since the direction of the contrasts was largely meaningless for the cross-sectional contrasts (and we later combined the positive and negative ratings as long as they were concordant) it may have resulted in better concordance if we had asked more simply for their opinion on the size of the difference between the groups alone.

Although traditional meta-analyses seek to combine all available information it is key to note that here it is actually better to only include the best possible information from the reviews, otherwise contrasts could be placed in the wrong size class for analysis. By using the subset of contrasts where experts were in broad agreement and had reviewed in line with the actual scores from the paper, we aimed to ensure the contrasts contributed to the correct size class estimate. Our analyses showed that if all contrasts were included regardless of quality, the estimates for the size classes did not show the same trend of increasing size from trivial through to large as would be expected.

A limitation of this work is that the decision on which contrasts to exclude from the analysis was made post hoc. There are however arguments against making a priori decisions on exclusion of studies in meta-analyses(99). Even in standard applications of meta-analysis there can be little evidence to base decision rules on at the start of the evidence synthesis. In this novel application of meta-analytic techniques, characteristics of contrasts affecting study quality were not clear a priori. Cooper *et al*(99) summarise by saying "Ironically, scholars who rely on a priori strategies may be excluding the evidence that may help them begin to establish the rules". This study was the first opportunity to conduct a thorough investigation of the meaning of quality evidence in this new application of meta-analytic techniques and it was important to initially include all available evidence.

It is possible that post hoc exclusion of contrasts could bias the results. In order to minimise possible bias, exclusions were based simply on whether the contrast could have been placed in the wrong size class for analysis. Although the quality assessment carried out could have led to more specific rules for exclusion (e.g. based on study characteristics), I felt that only excluding contrasts with a high probability of skewing the results minimised the risk of bias. Sensitivity analysis with the full set of contrasts was carried out in order to compare results with the chosen analysis dataset. This analysis confirmed that, when all contrasts were included, there was evidence of contrasts being placed in inappropriate size classes and leading to estimates that were clearly incorrect, such as trivial estimates above zero and medium estimates below zero.

For future studies there are two ways of minimising the risk of bias due to excluding contrasts. Firstly, one could try to minimise the disagreements between experts by using a consensus procedure. Secondly, the disagreements between experts and the actual QOL scores could be avoided by simplifying the review scale so experts judge only the size of the difference rather than the direction as well. This would avoid the kind of errors that led to the necessity for the high proportion of exclusions.

Despite using a subset of the better quality contrasts, our results still showed some overlap between the size class estimates. For the cross-sectional analysis there was overlap between the small/medium estimates for emotional, social, cognitive, constipation and dyspnoea subscales. Emotional functioning was not included in the cross-sectional guidelines since the estimates did not show a trend for increasing estimates as the size class increased. (Although I also found this for some of the longitudinal subscales they could generally be explained by small samples sizes.) This may be an indication that the emotional functioning subscale is hard for experts to predict and an area where the use of patient opinions may be more informative. However, the subscale also showed one of the smallest ranges of reported mean differences despite the wide range of clinical anchors, so it may be that this subscale is less responsive to change than would be expected. The cognitive function subscale similarly showed a narrower range of observed mean differences than the other subscales. The CF subscale also had a relatively high average baseline value for the longitudinal comparisons leaving little room for improvements over time. It is likely that the larger changes in these subscales would arise between groups receiving psycho-social related anchors/interventions which are not common in the literature. When planning a study it is likely that these subscales are appropriate as primary endpoints only in such interventions, as they are unlikely to be changed systematically in other situations.

For the longitudinal analysis there was a much bigger problem with the confidence intervals across size classes overlapping, with only one subscale having clear distinct estimates for each size class. We would conclude that the differences over time were less predictable for the experts, for a number of possible reasons. Health professionals and researchers such as those used in our expert panel may be more used to looking at comparisons between groups (e.g. from reported QOL in clinical trial groups or in their clinical practice) rather than following a patient's QOL over time, therefore may have found decisions about the size of difference in cross-

sectional comparisons easier. Qualitative interviews with the panel members did not support this though with an equal number of the experts reporting either cross-sectional or longitudinal comparisons as harder to judge.

It is possible that the anchors involved in judging between groups of patients (e.g. treatments, performance status etc.) could be more easily related to changes in QOL than the more general anchor of time alone. In considering a group of patients over time it may be there are too many factors changing (which may be known or unknown from a published paper) for an assessment of the average change to be accurate. Changes such as the nature of dropouts over time, response shift (psychological adaptation to changing health status) in participants filling in the questionnaires, on-going toxicity and further treatments outside of the study patients may receive could all effect the group's QOL but may not be reported(1). However, here we found that concordance between reviewers improved with increased dropout and they also reviewed with more certainty in their ratings. This may be an indication that because experts are familiar with the nature of dropouts generally in cancer studies (i.e. the patients with poorer QOL dropout) they can in fact predict more easily changes in the presence of the informative dropout.

9.4 Overall conclusion

This novel methodology resulted in guidelines for the QLQ-C30 which can now be used to aid the design and interpretation of clinical trials. The guidelines were shown to be robust to changes in the methodology and can reliably inform the meaning of changes in QOL scores from the QLQ-C30.

This research highlights the need for careful consideration at the design stage of a study of the expected size of QOL differences. Our new guidelines can be used to guide the sample size calculation and clinical interpretation of studies which compare the QOL of groups of patients or assess changes in QOL of groups over time. For the first time researchers can use guidelines that were developed specifically for the QLQ-C30 using previously observed change scores from the questionnaire. They can also interpret or base sample size calculations on individual subscales. When considering changes over time the guidelines reflect the fact that the meaning of improvements and deteriorations in QOL are not necessarily the same.

The methodology could now be applied to other QOL instruments with a sufficient evidence-base already established. I would recommend using a panel containing both

experts and patients along with a formal consensus process in order to obtain the opinions on size class. Using our investigation of factors affecting the quality of the contrasts, stricter inclusion criteria could now be employed, reducing the workload for the panel and the need for post hoc exclusions.

References

(1) King MT, Stockler MR, Cella DF, Osoba D, Eton DT, Thompson J, et al. Meta-analysis provides evidence-based effect sizes for a cancer-specific quality-of-life questionnaire, the FACT-G. J Clin Epidemiol 2010 Mar;63(3):270-81.

(2) King MT, Cella D, Osoba D, Stockler MR, Eton D, Thompson J, et al. Meta-analysis provides evidence-based interpretation guidelines for the clinical significance of mean differences for the FACT-G, a cancer-specific quality of life questionnaire. Patient Reported Outcome Measures 2010 Sep 22;1:119-26.

(3) Cocks K, King MT, Velikova G, Fayers PM, Brown JM. Quality, interpretation and presentation of European Organisation for Research and Treatment of Cancer quality of life questionnaire core 30 data in randomised controlled trials. Eur J Cancer 2008;44:1793-8.

(4) Cocks K, King MT, Velikova G, Martyn St-James M, Fayers PM, Brown JM. Evidence-Based Guidelines for Determination of Sample Size and Interpretation of the European Organisation for the Research and Treatment of Cancer Quality of Life Questionnaire Core 30. Journal of Clinical Oncology 2011 Jan 1;29(1):89-96.

(5) Anon. (2011). EORTC group for research into Quality of Life. Accessed on 28-6-2011 from http://groups.eortc.be/qol/

(6) Ferrans CE, Zerwic JJ, Wilbur JE, Larson JL. Conceptual model of health-related quality of life. Journal of Nursing Scholarship 2005;37(4):336-42.

(7) Ferrans CE. Development of a quality of life index for patients with cancer. Oncology Nursing Forum 1990;17(3):15-9.

(8) 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 1;11(3):570-9.

(9) Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, et al. The European Organisation for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. Journal of the National Cancer Institute 1993;85(5):365-76.

(10) Lydick E, Epstein RS. Interpretation of quality of life changes. Qual Life Res 1993;2(3):221-6.

(11) Kerlinger FN. Foundations of Behavioral Research. 2 ed. New York: 1973.

(12) Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR, Clinical Significance Consensus Meeting G. Methods to explain the clinical significance of health status measures. Mayo Clin Proc 2002;77(4):371-83.

(13) Lydick E. Approaches to the interpretation of quality-of-life scales.[comment]. Medical Care 2000;38(9 Suppl):II180-II183.

(14) Marquis P, Chassany O, Abetz L. A comprehensive strategy for the interpretation of quality-of-life data based on existing methods.[see comment]. Value Health 2004;7(1):93-104.

(15) Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. Journal of Clinical Epidemiology 2008 Feb;61(2):102-9.

(16) King MT. A point of minimal important difference (MID): a critique of terminology and methods . Expert Review of Pharmacoeconomics & Outcomes. In press 2011.

(17) Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials 1989;10(4):407-15.

(18) de Vet HC, Terwee CB, Ostelo RW, Beckerman H, Knol DL, Bouter LM. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. Health Qual Life Outcomes 2006;4:54.

(19) King MT. The interpretation of scores from the EORTC quality of life questionnaire QLQ-C30. Qual Life Res 1996;5(6):555-67.

(20) King MT. Cohen confirmed? Empirical effect sizes for the QLQ-C30. Qual Life Res 2001;10:278.

(21) Wyrwich KW, Bullinger M, Aaronson N, Hays RD, Patrick DL, Symonds T. Estimating clinically significant differences in quality of life outcomes. Quality of Life Research 2005 Mar 1;14(2):285-95.

(22) Juniper EF, Guyatt GH, Willan A, Griffith LE. Determining a minimal important change in a disease-specific Quality of Life Questionnaire. J Clin Epidemiol 1994;47(1):81-7.

(23) Kamper SJ, Maher CG, Mackay G. Global rating of change scales: a review of strengths and weaknesses and considerations for design. J Man Manip Ther 2009;17(3):163-70.

(24) Herrmann D. Reporting current, past, and changed health status. What we know about distortion. Med Care 1995 Apr;33(4 Suppl):AS89-AS94.

(25) Schwartz CE, Sprangers MA. Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. Soc Sci Med 1999 Jun;48(11):1531-48.

(26) Guyatt GH, Norman GR, Juniper EF, Griffith LE. A critical look at transition ratings. J Clin Epidemiol 2002 Sep;55(9):900-8.

(27) Schmitt J, Di Fabio RP. The validity of prospective and retrospective global change criterion measures. Arch Phys Med Rehabil 2005 Dec;86(12):2270-6.

(28) Redelmeier DA, Guyatt GH, Goldstein RS. On the debate over methods for estimating the clinically important difference.[comment]. J Clin Epidemiol 1996;49(11):1223-4.

(29) Redelmeier DA, Guyatt GH, Goldstein RS. Assessing the minimal important difference in symptoms: a comparison of two techniques.[see comment]. J Clin Epidemiol 1996;49(11):1215-9.

(30) Wyrwich KW. Minimal important difference thresholds and the standard error of measurement: is there a connection? Journal of Biopharmaceutical Statistics 2004;14(1):97-110.

(31) Cohen J. Statistical power analysis for the behavioral sciences (rev. ed.). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc; 1977.

(32) Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation.[see comment]. Medical Care 2003;41(5):582-92.

(33) Wright JG. Interpreting health-related quality of life scores: the simple rule of seven may not be so simple.[see comment][comment]. Medical Care 2003;41(5):597-8.

(34) Beaton DE. Simple as possible? Or too simple? Possible limits to the universality of the one half standard deviation.[see comment][comment]. Medical Care 2003;41(5):593-6.

(35) Wyrwich KW, Tierney WM, Wolinsky FD. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. J Clin Epidemiol 1999;52(9):861-73.

(36) Cella D, Eton DT, Lai JS, Peterman AH, Merkel DE. Combining anchor and distribution-based methods to derive minimal clinically important differences on the Functional Assessment of Cancer Therapy (FACT) anemia and fatigue scales. J Pain Symptom Manage 2002;24(6):547-61.

(37) Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. Journal of Consulting & Clinical Psychology 1991 Feb;59(1):12-9.

(38) Crosby RD, Kolotkin RL, Williams GR. An integrated method to determine meaningful changes in health-related quality of life. J Clin Epidemiol 2004;57(11):1153-60.

(39) Yost KJ, Eton DT. Combining Distribution- and Anchor-Based Approaches to Determine Minimally Important Differences. Evaluation & the Health Professions 2005 Jun 1;28(2):172-91.

(40) Denzin NK. The research act: A theoretical introduction to sociological methods. New York: McGraw-Hill; 1978.

(41) Leidy NK, Wyrwich KW. Bridging the gap: using triangulation methodology to estimate minimal clinically important differences (MCIDs). COPD 2005 Mar;2(1):157-65.

(42) Revicki DA, Erickson PA, Sloan JA, Dueck A, Guess H, Santanello NC. Interpreting and Reporting Results Based on Patient-Reported Outcomes. Value in Health 2011 Nov;10(Supplement 2):S116-S124.

(43) de Vet H, Ostelo R, Terwee C, van der Roer N, Knol D, Beckerman H, et al. Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. Quality of Life Research 2007 Feb 1;16(1):131-42.

(44) Fayers PM. Interpreting quality of life data: population-based reference data for the EORTC QLQ-C30.[comment]. Eur J Cancer 2001;37(11):1331-4.

(45) Schwarz R, Hinz A. Reference data for the quality of life questionnaire EORTC QLQ-C30 in the general German population. Eur J Cancer 2001 Jul;37(11):1345-51.

(46) Klee M, Groenvold M, Machin D. Quality of life of Danish women: population-based norms of the EORTC QLQ-C30. Qual Life Res 1997;6(1):27-34.

(47) Hjermstad MJ, Fayers PM, Bjordal K, Kaasa S. Health-related quality of life in the general Norwegian population assessed by the European Organization for Research and Treatment of Cancer Core Quality-of-Life Questionnaire: the QLQ-C30 (+ 3). J Clin Oncol 1998;16(3):1188-96.

(48) Michelson H, Bolund C, Nilsson B, Brandberg Y. Health-related quality of life measured by the EORTC QLQ-C30--reference values from a large sample of Swedish population. Acta Oncol 2000;39(4):477-84.

(49) Osoba D, Rodrigues G, Myles J, Zee B, Pater J. Interpreting the significance of changes in health-related quality-of-life scores. J Clin Oncol 1998;16(1):139-44.

(50) Guyatt GH, Schunemann HJ. How can quality of life researchers make their work more useful to health workers and their patients? Qual Life Res 2007;16:1097-105.

(51) Efficace F, Bottomley A, van Andel G. Health-related quality of life in prostate carcinoma patients: a systematic review of randomized controlled trials. Cancer 2003;97:377-88.

(52) Goodwin PJ, Black JT, Bordeleau LJ, Ganz PA. Health-related quality-of-life measurement in randomized clinical trials in breast cancer - taking stock. Journal of the National Cancer Institute 2003;95(4):263-81.

(53) Blazeby JM, Avery K, Sprangers M, Pikhart H, Fayers PM, Donovan J. Health-related quality of life measurement in randomized clinical trials in surgical oncology. J Clin Oncol 2006;24(19):3178-86.

(54) Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med 2001;134(8):663-94.

(55) Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. Ann Intern Med 2001;134(8):657-62.

(56) Efficace F, Bottomley A, Osoba D, Gotay C, Flechtner H, D'haese S, et al. Beyond the development of health-related quality-of-life (HRQOL) measures: a checklist for evaluating HRQOL outcomes in cancer clinical trials--does HRQOL evaluation in prostate cancer research inform clinical decision making? Journal of Clinical Oncology 21(18):3502-11, 2003 Sep 15.

(57) Osoba D, Bezjak A, Brundage M, Zee B, Tu D, Pater J, et al. Analysis and interpretation of health-related quality-of-life data from clinical trials: basic approach of The National Cancer Institute of Canada Clinical Trials Group. Eur J Cancer 2005;41(2):280-7.

(58) Guyatt GH, Juniper EF, Walter SD, Griffith LE, Goldstein RS. Interpreting treatment effects in randomised trials.[see comment]. BMJ 1998;316(7132):690-3.

(59) Osoba D, King MT. Meaningful differences. In: Fayers PM, Hays RD, editors. Assessing quality of life in clinical trials. Oxford University Press; 2005.

(60) Anon. (2011). EORTC group for research into Quality of Life: Bibliography. Accessed on 11-7-2011 from http://groups.eortc.be/qol/documentation_bibliography.htm

(61) Sloan JA. Asking the obvious questions regarding patient burden. J Clin Oncol 2002;20:4-6.

(62) Lee R.W., McQuellon R.P., Case L.D., DeGuzman A.F., McCullough D.L. Early quality of life assessment in men treated with permanent source interstitial brachytherapy for clinically localized prostate cancer. J Urol 1999;162:403-6.

(63) Glimelius B, Hoffman K, Pahlman L. Monitoring palliative chemotherapy in advanced gastrointestinal cancer using serial tissue polypeptide antigen specific (TPS) measurements. Acta Oncologica 1996;35:141-8.

(64) Bottomley A, Therasse P. Quality of life in patients undergoing systemic therapy for advanced breast cancer. Lancet Oncology 2002;3:620-8.

(65) Efficace F, Bottomley A, Vanvoorden V, Blazeby JM. Methodological issues in assessing health-related quality of life of colorectal cancer patients in randomized controlled trials. Eur J Cancer 2004;40:187-97.

(66) Bottomley A, Efficace F, Thomas R, Vanvoorden V, Ahmedzai S. Health-related quality of life in non small-cell lung cancer: methodologic issues in randomized clinical trials. J Clin Oncol 2003;21:2982-92.

(67) Efficace F, Hornebar M, Lejeune S, Van Dam F, Leering S, Rottman M, et al. Methodological quality of patient-reported outcome research was low in complementary and alternative medicine in oncology. J Clin Epidemiol 2006;59:1257-65.

(68) Cocks, K. (2011). EBIG Expert Review Manual. Accessed on 11-7-2011 from http://ctru.leeds.ac.uk/ebig

(69) Alderson P., Green S., Higgins J.P.T. Cochrane Reviewers' Handbook 4.2.2. In: The Cochrane Library, Issue 1, 2004. ed. Chichester, UK: John Wiley & Sons; 2006.

(70) Follmann D, Elliott P, Suh I, Cutler J. Variance imputation for overviews of clinical trials with continuous response. J Clin Epidemiol 1992;45(7):769-73.

(71) Lipsey MW, Wilson DB. Practical Meta-Analysis. Sage Publications Inc.; 2001.

(72) Furukawa TA, Barbui C, Cipriani A, Brambilla P, Watanabe N. Imputing missing standard deviations in meta-analyses can provide accurate results. J Clin Epidemiol 2006 Jan;59(1):7-10.

(73) Fayers PM, Aaronson NK, Bjordal K, Sullivan M. EORTC QLQ-C30 Scoring manual. Brussels: EORTC Quality of Life Study Group; 1995.

(74) Hedges LV, Olkin I. Statistical Methods for Meta-Analysis. Academic Press, Inc.; 1985.

(75) Lipsey MW, Wilson DB. Practical Meta-Analysis. Sage Publications Inc.; 2001.

(76) Follmann D, Elliott P, Suh I, Cutler J. Variance imputation for overviews of clinical trials with continuous response. J Clin Epidemiol 1992;45(7):769-73.

(77) Higgins JPT, Green S.(editors). Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.2. 2009.

(78) Hedges LV. Distribution theory for Glass's estimator of effect size and related estimators. Journal of Educational Statistics 1981;6(2):107-28.

(79) Wang MC, Bushman BJ. Integrating results through meta-analytic review using SAS software. Cary, NC: SAS Institute Inc.; 1999.

(80) Olivo SA, Macedo LG, Gadotti IC, Fuentes J, Stanton T, Magee DJ. Scales to assess the quality of randomized controlled trials: a systematic review. Physical Therapy 2008;88:156-72.

(81) Tritchler D. Modelling study quality in meta-analysis. Statistics in Medicine 1999;18:2135-45.

(82) Glover, J., Izzo, D., Odato, K., and Wang, L. (2011). EBM Page Generator. Accessed on 28-6-2011 from www.ebmpyramid.org

(83) Tastle WJ, Wierman MJ. An information theoretic measure for the evaluation of ordinal scale data. Behavior Research Methods 2006;38:487-94.

(84) Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin 1979;86(2):420-8.

(85) Fleiss JL. The design and analysis of clinical experiments. New York: Wiley; 1986.

(86) Cella D, Hahn EA, Dineen K. Meaningful change in cancer-specific quality of life scores: differences between improvement and worsening. Qual Life Res 2002;11(3):207-21.

(87) Hancock B. Trent Focus for Research and Development in Primary Health Care: An Introduction to Qualitative Research. Trent Focus; 1998.

(88) Bland JM, Altman DG. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. Computers in Biology and Medicine 1990;20:337-40.

(89) Redelmeier DA, Goldstein RS, Min ST, Hyland RH. Spirometry and dyspnea in patients with COPD. When small differences mean little. Chest 1996 May;109(5):1163-8.

(90) Ringash J, O'Sullivan B, Bezjak A, Redelmeier DA. Interpreting clinically significant changes in patient-reported outcomes. Cancer 2007 Jul 1;110(1):196-202.

(91) Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups.

(92) Tourangeau R. Cognitive sciences and survey methods. In: Jabine T, Straf M, Tanur J., Tourangeau R, editors. Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines.Washington,DC: National Academy Press; 1984. p. 73-100.

(93) Brundage M, Feldman-Stewart D, Leis A, Bezjak A, Degner L, Velji K, et al. Communicating quality of life information to cancer patients: a study of six presentation formats. Journal of Clinical Oncology 2005 Oct 1;23(28):6949-56.

(94) Brundage M, Leis A, Bezjak A, Feldman-Stewart D, Degner L, Velji K, et al. Cancer patients' preferences for communicating clinical trial quality of life information: a qualitative study. Quality of Life Research 2003 Jun;12(4):395-404.

(95) The PROMIS Network. (2011). PROMIS: Dynamic Tools to Measure Health Outcomes from the Patient Perspective. Accessed on 11-7-2011 from http://www.nihpromis.org/

(96) Bryman A, Burgess RG. Analyzing qualitative data. 1 ed. Routledge; 1994.

(97) Sprangers MA. Response-shift bias: a challenge to the assessment of patients' quality of life in cancer clinical trials. Cancer Treatment Reviews 1996;22(Suppl A):55-62.

(98) Stephens RJ, Hopwood P, Girling DJ, Machin D. Randomized trials with quality of life endpoints: Are doctors' ratings of patients' physical symptoms interchangeable with patients' self-ratings? Quality of Life Research 1997;6:225-36.

(99) Cooper HM, Hedges LV, Valentine JC. The handbook of research synthesis and meta-analysis. Russell Sage Foundation; 2009.

Appendix I Example coversheet

Ref ID = 295	Summary details of paper
Research question	To measure the effects of adjuvant chemotherapy on physical function and HRQOL
Design	Cohort/descriptive study
Country	Canada
Disease	Breast
Disease extent	Mixed
Males:Females	0%:100%
Study size	68
Average age	60

Details of comparisons

Between-group	Anchor	Comparison	Timepoint description	Data source
X1	Age	Young group vs Older group	Prior to chemotherapy	Table 2 p1746
X2	Age	Young group vs Older group	Third cycle of chemotherapy	Table 2 p1746
X3	Age	Young group vs Older group	Completion of chemotherapy	Table 2 p1746
X4	Age	Young group vs Older group	Six months post-chemotherapy	Table 2 p1746

Longitudinal	Anchor	Comparison	Group description	Data source
L1	Treatment	Prior to chemotherapy vs Third cycle of chemotherapy	Young group	Table 2 p1746
L2	Treatment	Prior to chemotherapy vs Completion of chemotherapy	Young group	Table 2 p1746
L3	Treatment	Prior to chemotherapy vs Six months post-chemotherapy	Young group	Table 2 p1746
L4	Treatment	Prior to chemotherapy vs Third cycle of chemotherapy	Older group	Table 2 p1746
L5	Treatment	Prior to chemotherapy vs Completion of chemotherapy	Older group	Table 2 p1746
L6	Treatment	Prior to chemotherapy vs Six months post-chemotherapy	Older group	Table 2 p1746

For between-group comparisons, consider each table. For each possible contrast, and for each aspect of QOL, how do you expect the mean QOL score of the 1st group to differ from the mean QOL score of the 2nd group? For longitudinal comparisons, consider each table. For each possible contrast, and for each aspect of QOL, how do you expect the mean QOL score to change over time?

The following scale is used in the tables:-

-3	-2	-1	0	1	2	3
Much worse	Moderately worse	A little worse	Much the same	A little better	Moderately better	Much better

Between Group Comparisons

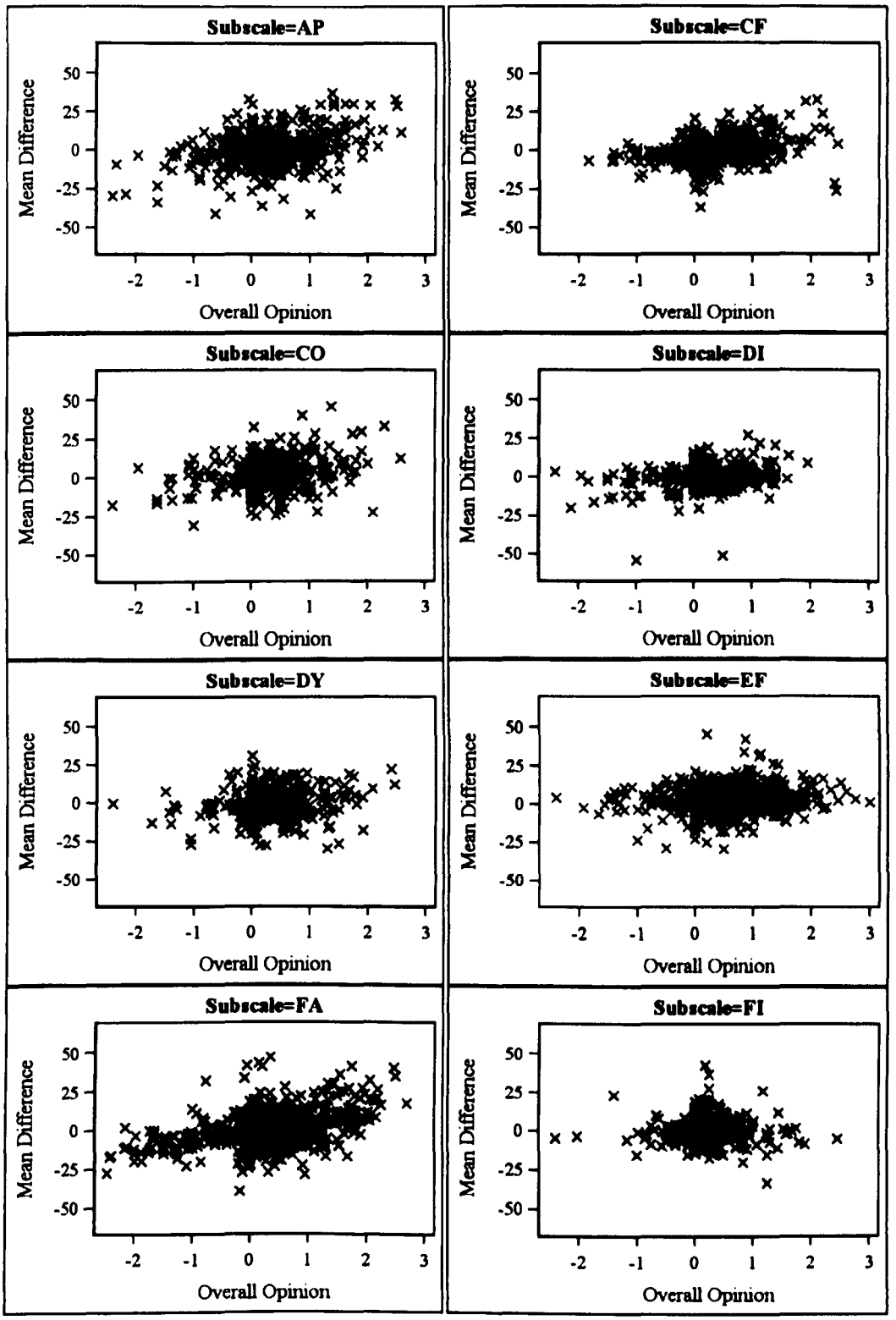
Anchor: Age G1=Young G2=Older	Comparison X1 Pre-chemotherapy G1 (n=43) vs G2 (n=16)							Comparison X2 After cycle 3 G1 (n=37) vs G2 (n=18)							Comparison X3 Completion of chemo G1 (n=29) vs G2 (n=18)							Comparison X4 6 months post chemo G1 (n=29) vs G2 (n=15)						
	G1 worse				G1 better			G1 worse				G1 better			G1 worse				G1 better			G1 worse				G1 better		
	-3	-2	-1	0	1	2	3	-3	-2	-1	0	1	2	3	-3	-2	-1	0	1	2	3	-3	-2	-1	0	1	2	3
Physical functioning																												
Role functioning																												
Emotional functioning																												
Social functioning																												
Cognitive functioning																												
Nausea and vomiting																												
Fatigue																												
Pain																												
Global health status																												
Dyspnoea																												
Sleep																												
Appetite loss																												
Constipation																												
Diarrhoea																												
Financial Difficulties																												

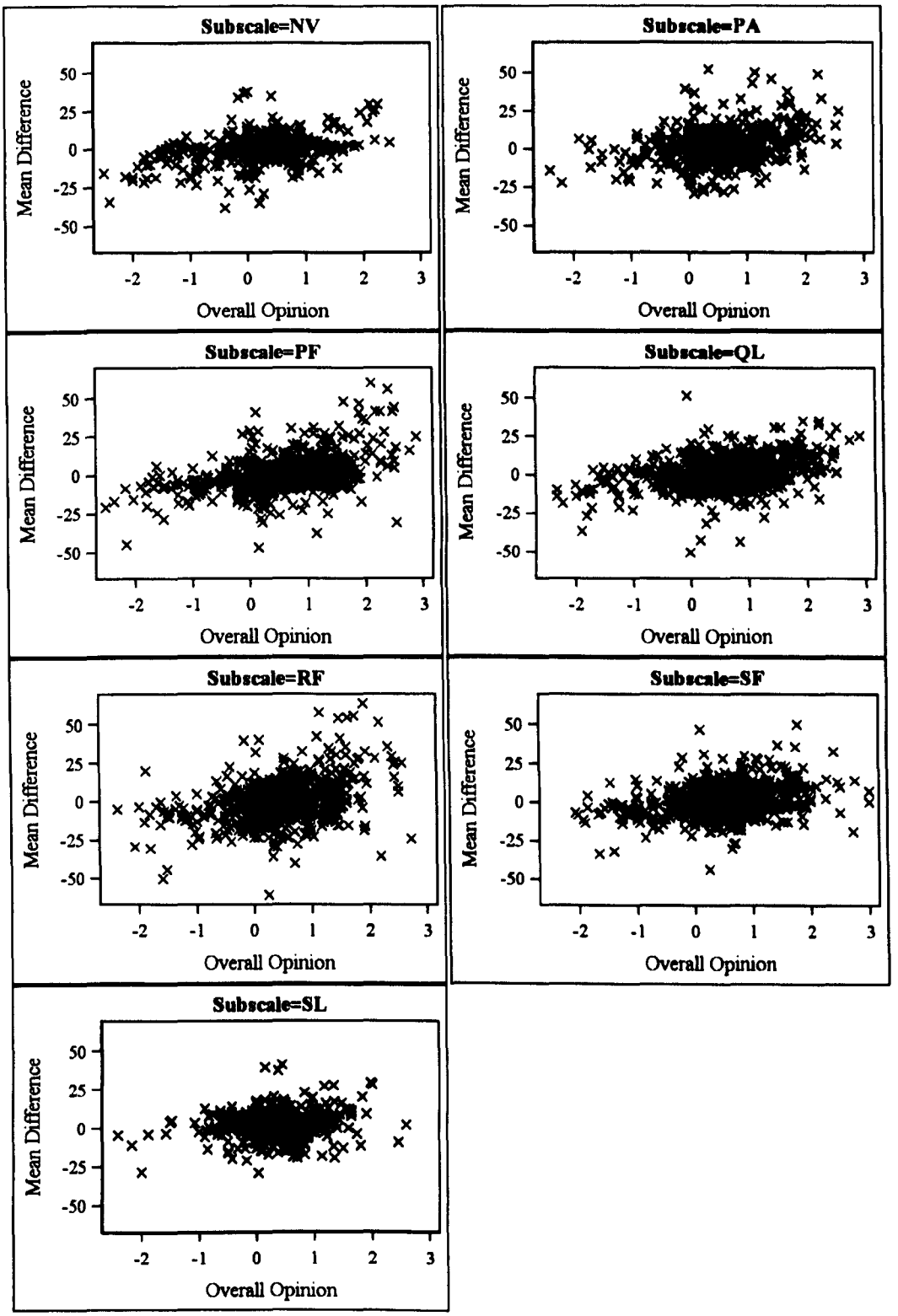
Longitudinal comparisons

Anchor: Treatment T1= Pre-chemo T2=3rd cycle of chemo T3= Completion of chemotherapy T4=6 months post chemo	Comparison L1 Group: Young T1 (n=43) vs T2 (n=37)							Comparison L2 Group: Young T1 (n=43) vs T3 (n=29)							Comparison L3 Group: Young T1 (n=43) vs T4 (n=29)						
	T2 worse				T2 better			T3 worse				T3 better			T4 worse				T4 better		
	-3	-2	-1	0	1	2	3	-3	-2	-1	0	1	2	3	-3	-2	-1	0	1	2	3
Physical functioning																					
Role functioning																					
Emotional functioning																					
Social functioning																					
Cognitive functioning																					
Nausea and vomiting																					
Fatigue																					
Pain																					
Global health status																					
Dyspnoea																					
Sleep																					
Appetite loss																					
Constipation																					
Diarrhoea																					
Financial Difficulties																					

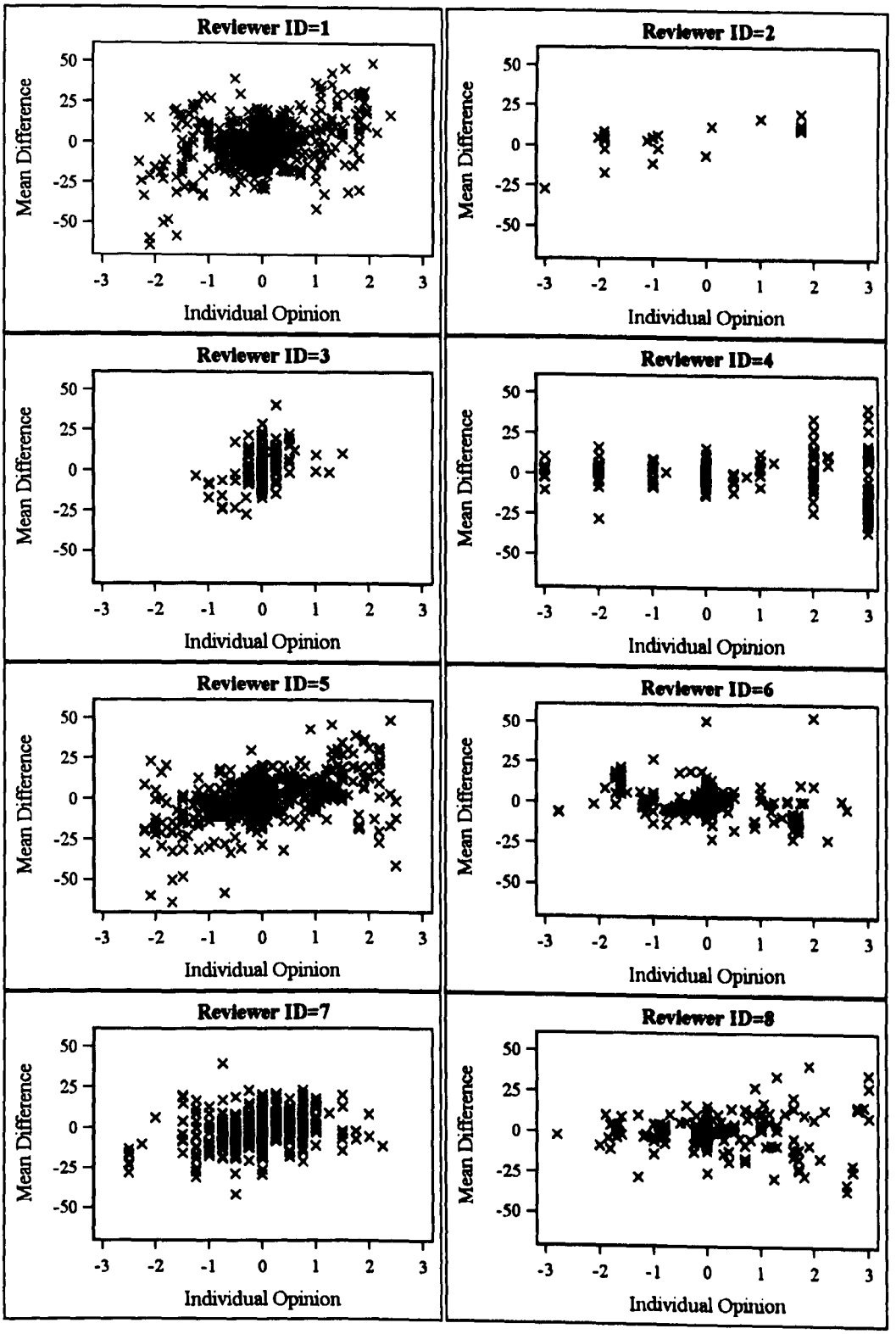
Anchor: Treatment	Comparison L4 Group: Older T1 (n=16) vs T2 (n=18)							Comparison L5 Group: Older T1 (n=16) vs T3 (n=18)							Comparison L6 Group: Older T1 (n=16) vs T4 (n=15)						
	T2 worse				T2 better			T3 worse				T3 better			T4 worse				T4 better		
	-3	-2	-1	0	1	2	3	-3	-2	-1	0	1	2	3	-3	-2	-1	0	1	2	3
Physical functioning																					
Role functioning																					
Emotional functioning																					
Social functioning																					
Cognitive functioning																					
Nausea and vomiting																					
Fatigue																					
Pain																					
Global health status																					
Dyspnoea																					
Sleep																					
Appetite loss																					
Constipation																					
Diarrhoea																					
Financial Difficulties																					

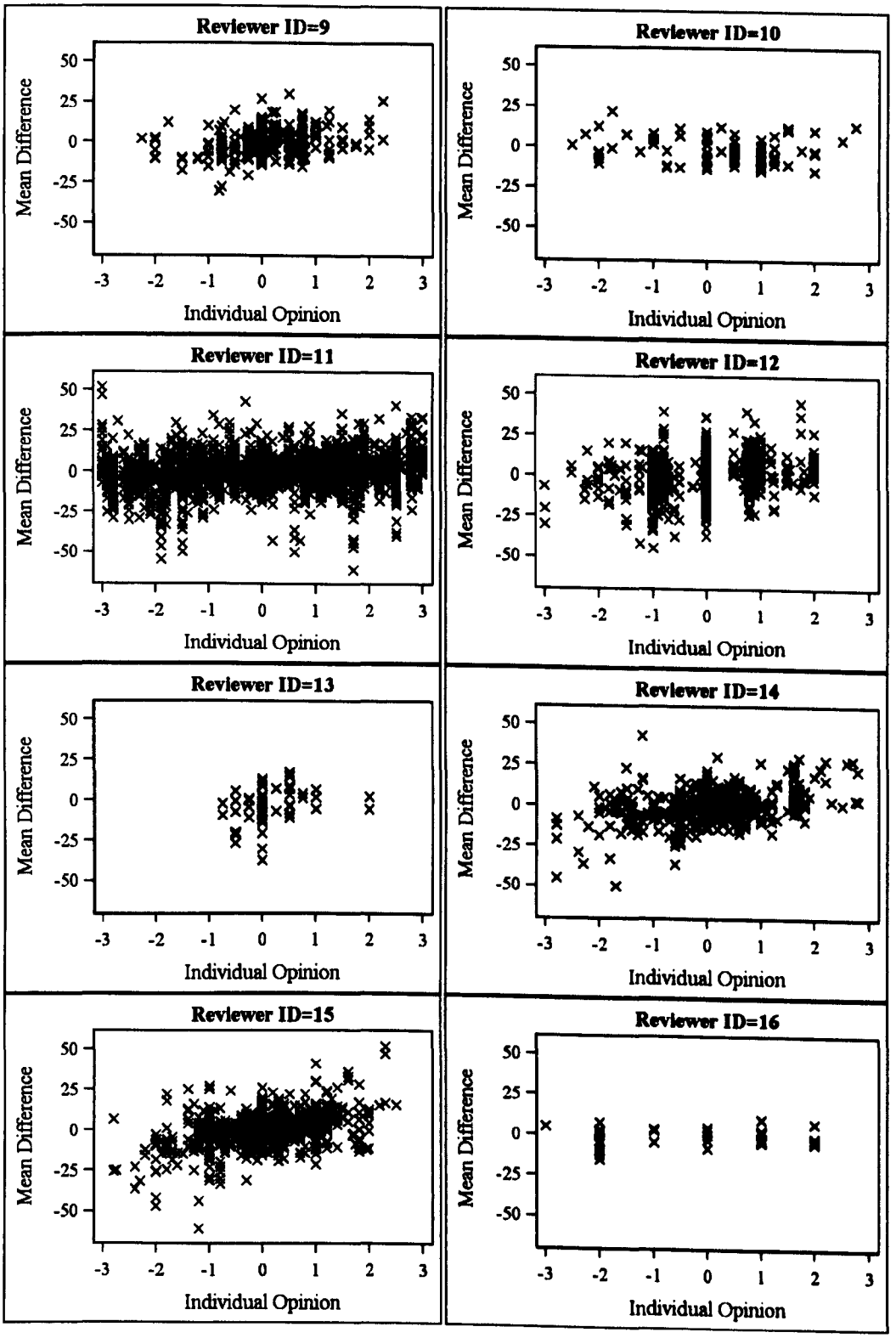
Appendix II Mean difference versus overall opinion by subscale

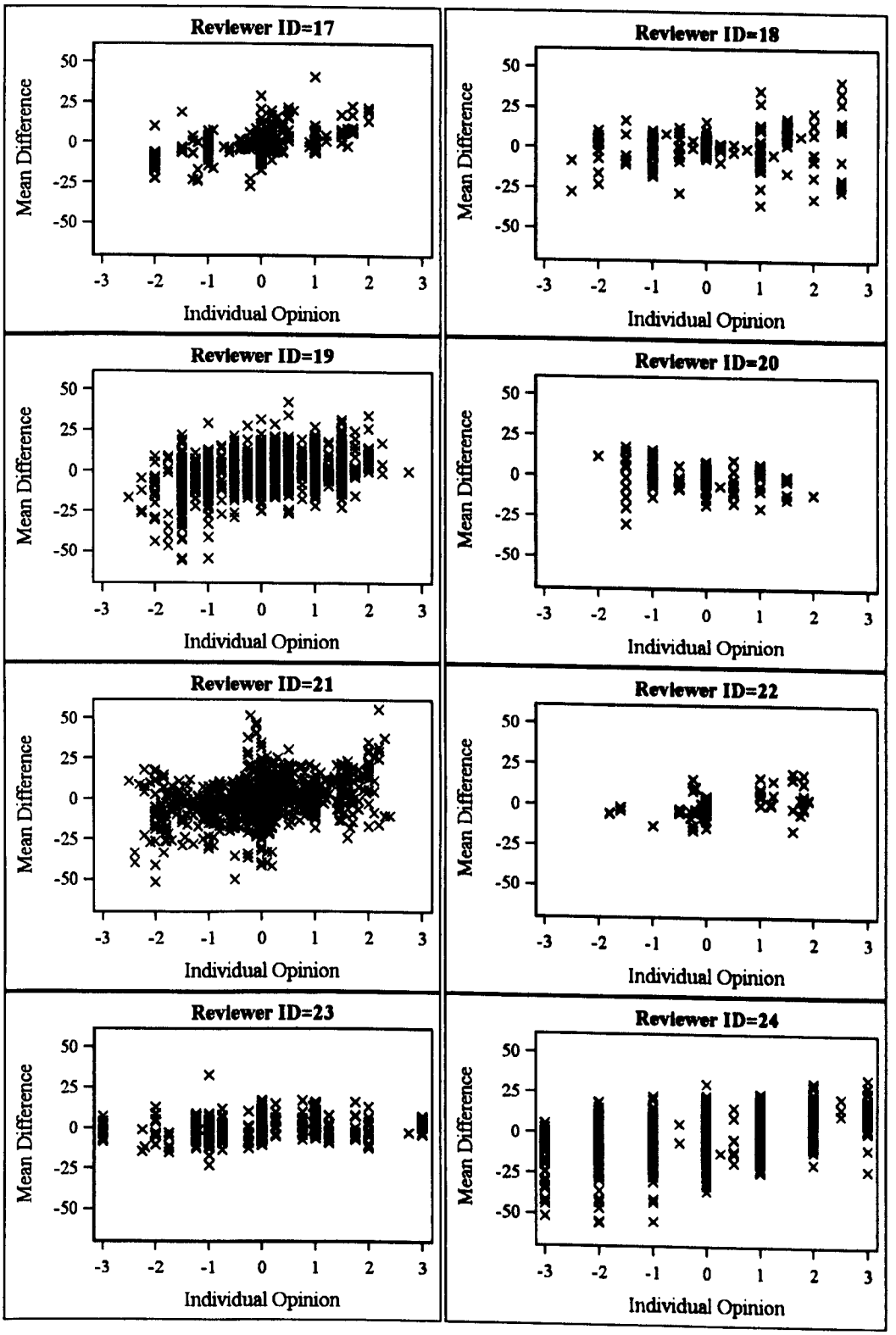


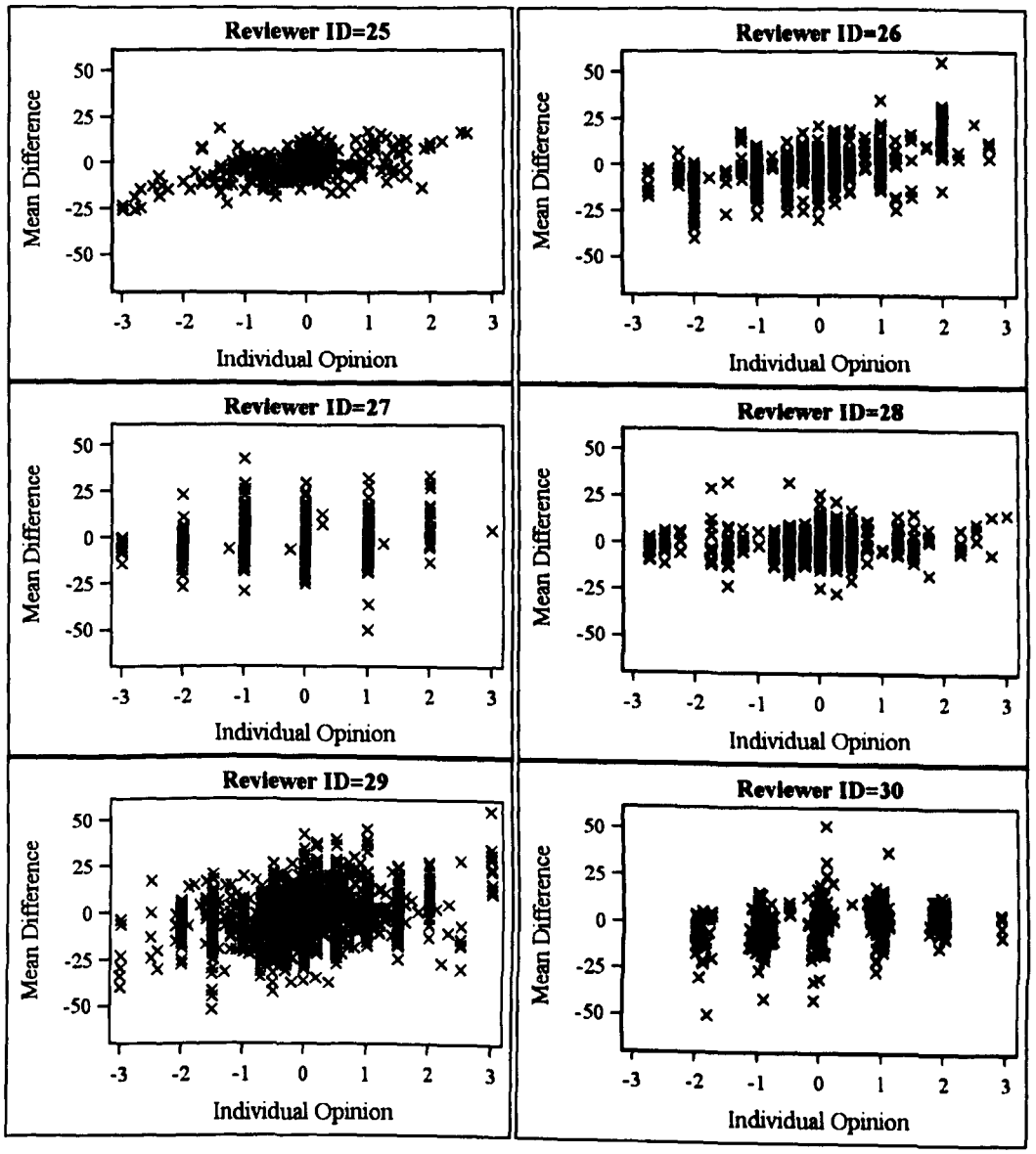

Subscale=NV
Subscale=PA
Subscale=PF
Subscale=QL
Subscale=RF
Subscale=SF
Subscale=SL
Mean Difference
Overall Opinion
50
25
0
-25
-50
-2
-1
0
1
2
3

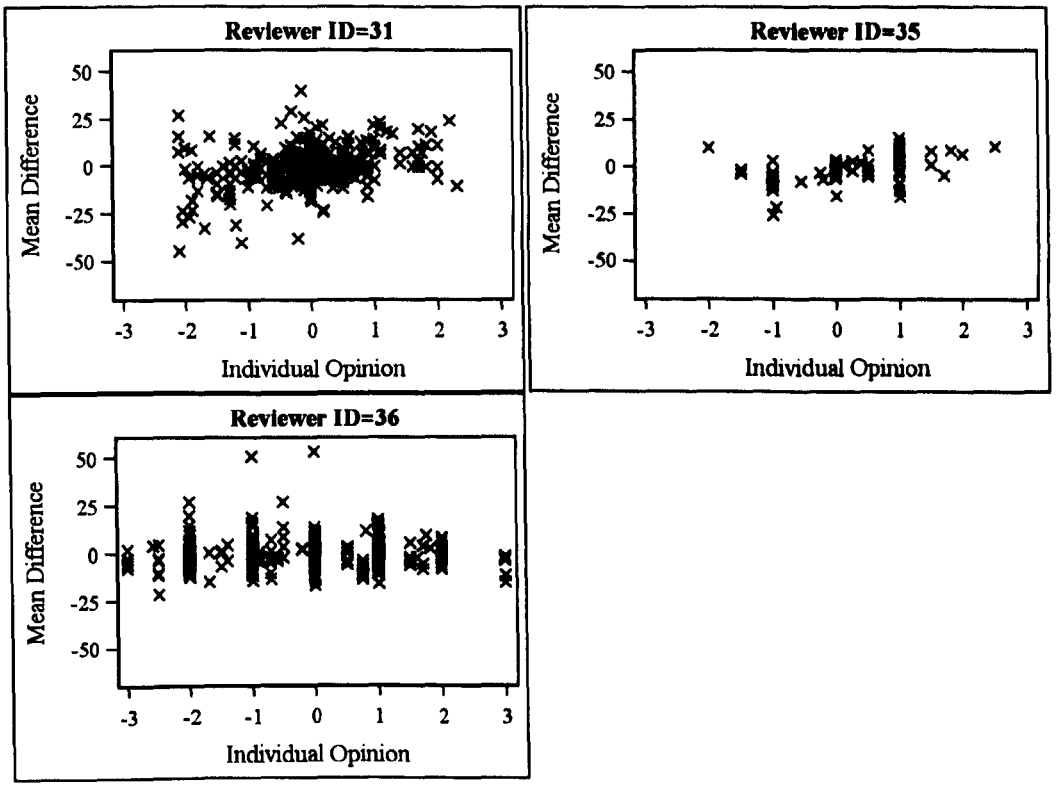
Appendix III Mean difference versus individual reviewer opinions




Reviewer ID=9
Reviewer ID=10
Reviewer ID=11
Reviewer ID=12
Reviewer ID=13
Reviewer ID=14
Reviewer ID=15
Reviewer ID=16
Mean Difference
Individual Opinion
50
25
0
-25
-50
-3
-2
-1
0
1
2
3


Reviewer ID=17
Reviewer ID=18
Reviewer ID=19
Reviewer ID=20
Reviewer ID=21
Reviewer ID=22
Reviewer ID=23
Reviewer ID=24
Mean Difference
Individual Opinion
50
25
0
-25
-50
-3
-2
-1
0
1
2
3


Reviewer ID=25
Reviewer ID=26
Reviewer ID=27
Reviewer ID=28
Reviewer ID=29
Reviewer ID=30
Mean Difference
Individual Opinion
50
25
0
-25
-50
-3
-2
-1
0
1
2
3


Reviewer ID=31
Mean Difference
50
25
0
-25
-50
-3
-2
-1
0
1
2
3
Individual Opinion
Reviewer ID=35
Mean Difference
50
25
0
-25
-50
-3
-2
-1
0
1
2
3
Individual Opinion
Reviewer ID=36
Mean Difference
50
25
0
-25
-50
-3
-2
-1
0
1
2
3
Individual Opinion

Appendix IV Patient information sheet (on hospital headed paper)

INFORMATION SHEET
Version 1.1
19 June 2007

Evidence-based interpretation guidelines for the QLQ-C30: A pilot patient sub-study

We would like to invite you to take part in a research study. Before you decide you need to understand why the research is being done and what it would involve for you. Please take time to read the following information carefully. Talk to others about the study if you wish.

(Part 1 tells you the purpose of this study and what will happen to you if you take part. Part 2 gives you more detailed information about the conduct of the study).

Ask us if there is anything that is not clear or if you would like more information. Take time to decide whether or not you wish to take part.

Part 1

What is the purpose of the study?

Cancer patients may often be asked to fill in quality of life questionnaires as part of their clinic appointment or in a clinical trial. While doctors and researchers are becoming more familiar with these questionnaires there is still an uncertainty over what differences in the answers are meaningful to patients. This study forms part of a PhD project aiming to provide guidelines for doctors and researchers using quality of life questionnaires with patients. The guidelines will suggest the size of differences in quality of life scores that matter to patients compared with differences that are not noticeable to a patient. This is an initial study involving interviews with patients like yourself, to work out the best way of asking patients about the size of changes in quality of life.

Why have I been invited?

Any patient attending the Medical Oncology breast cancer clinics at St James' Hospital in Leeds may be invited to participate.

Do I have to take part?

It is up to you to decide. We will describe the study and go through this information sheet, which we will then give to you. We will then ask you to sign a consent form to show you have agreed to take part. You are free to withdraw at any time, without giving a reason. This would not affect the standard of care you receive.

What will happen to me if I take part?

You will be contacted to arrange a suitable interview date and time. Kim Cocks will then meet with you to discuss various scenarios involving groups of breast cancer patients and their quality of life. The discussion should last around 1 hour and will be tape

recorded. The discussion will take place either in the breast cancer clinic at St James' hospital or at your home if you prefer. The tape recording will be used by researchers involved in the project to write notes on the discussion and will be stored in a locked cabinet. The tapes will be destroyed 10 years after the completion of the research.

We will ask you to complete a short checklist about yourself and your life circumstances. We also would like to ask for permission to use your hospital records to see what medication you are taking and what care you are receiving. The information will be confidential between you, the team looking after you and the researchers.

Expenses and payments
We anticipate that there will be no extra expenses for you as a result of taking part in this study, as interviews will be conducted while you are attending the hospital for an appointment where possible.

What will I have to do?
Attend an interview
Provide information on your life circumstances
Agree to researchers accessing your hospital records

What are the possible disadvantages and risks of taking part?
The study requires approximately one hour of your time. You will be asked to think about other breast cancer patients and their quality of life and there is a possibility you may find this distressing. The interview can be stopped at any point if you feel you do not want to continue. A referral can be made to your treating clinician if you are distressed by the content of the discussion.

What are the possible benefits of taking part?
We hope that the information we get from the interviews will lead to doctors and researchers having a better understanding of quality of life questionnaires and the meaning of a change in the questionnaire scores from a patient's perspective.

Will my taking part in the study be kept confidential?
Yes. We will follow ethical and legal practice and all information about you will be handled in confidence. The details are included in Part 2.

This completes part 1.

If the information in Part 1 has interested you and you are considering participation, please read the additional information in Part 2 before making any decision.

Part 2

What will happen if I don't want to carry on with the study?
You are free to change your mind at any point up to, during or following the interview. You will not be able to be identified in the study results but if you wish to withdraw any data already collected prior to publication of the results then arrangements can be made for the interview tape to be destroyed and your discussion excluded from the study.

Will my taking part in this study be kept confidential?
If you join the study, the information collected about you will be kept strictly confidential. Some parts of your medical records and the data collected for the study will be looked at by authorised persons from the University of Leeds. Mrs Kim Cocks will store the interview tapes in a locked cabinet. Tapes will be identified by study number only and any references to names removed during transcription. Identifiable data will only be accessed by the researchers. All researchers will have a duty of confidentiality to you as a research participant, under the provisions of the 1998 Data Protection Act. Data will be retained for 10 years and then disposed of securely.

Involvement of the General Practitioner/Family doctor (GP)
Your GP will not be notified of your participation in this study.

What will happen to the results of the research study?
Participants will not be identified in any report/publication. The study results will be published in a scientific journal.

Who is organising and funding the research?
Cancer Research UK is funding the research. The study is sponsored by the University of Leeds. The researchers and doctors are not being paid for inclusion of patients in this study.

Who has reviewed the study?
All research in the NHS is looked at by independent group of people, called a Research Ethics Committee to protect your safety, rights, wellbeing and dignity. This study has been reviewed and given favourable opinion by Leeds East Research Ethics Committee.

Further information and contact details
If you would like to discuss the study further please contact the Chief Investigator, Mrs Kim Cocks on 0113 3431475 or speak to Dr Galina Velikova on 0113 2064905.

Appendix V Scenario D for patient interviews

Scenario description given to patient

This study investigated the quality of life in post-menopausal women receiving anthracycline-based adjuvant chemotherapy. It compared the 'younger' women (less than 65) with the older women (65 or over). All women filled in the quality of life questionnaire before receiving the chemotherapy, just before cycle 3, three weeks after cycle 6 and at 6 and 12 months after chemotherapy.

Group 1: **Less than 65 years**

On average these patients were 55 years old, ages ranged from 31 to 64

Group 2: **65 years or over**

On average these patients were 70 years old, ages ranged from 65 to 80

Interview questions and visual aids

Interviewer: "First I would like you to consider whether these two groups would experience differences with respect to social functioning 6 months after chemotherapy. Here is a reminder of the 2 questions used to create a social functioning score and some examples of how the answers to the questions are changed into a score."

Questions describing social functioning

During the past week:

	Not at all	A little	Quite a bit	Very much
26. Has your physical condition or medical treatment interfered with your family life?	1	2	3	4
27. Has your physical condition or medical treatment interfered with your social activities?	1	2	3	4

Examples of possible scores

Score=100	Not at all	A little	Quite a bit	Very much
26. Has your physical condition or medical treatment interfered with your family life?	(1)	2	3	4
27. Has your physical condition or medical treatment interfered with your social activities?	(1)	2	3	4

Score=83	Not at all	A little	Quite a bit	Very much
26. Has your physical condition or medical treatment interfered with your family life?	(1)	2	3	4
27. Has your physical condition or medical treatment interfered with your social activities?	1	(2)	3	4

Interviewer: Do you think there would be a noticeable difference in social functioning measured using the two questions in this questionnaire between these groups?

If yes, do you think this difference would be a small, medium or large difference?

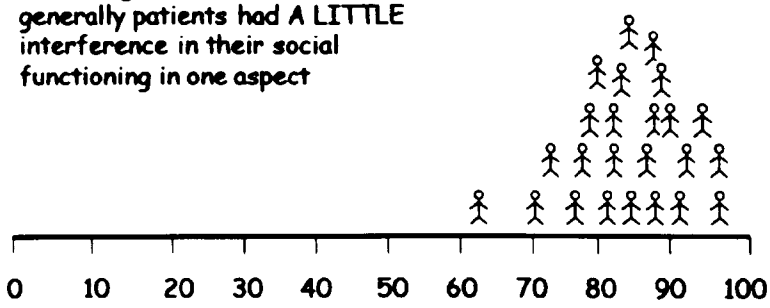
Interviewer: These pictures show what the two groups of women actually reported. Patients aged less than 65 scored on average 85 on the social functioning scale whereas patients 65 and over scored an average of 92 on the social functioning scale, does this change your opinion?

If yes, do you think this difference is a small, medium or large difference?

Less then 65 years at 6 months after chemotherapy

Average social functioning score = 85

An average score of 85 means generally patients had A LITTLE interference in their social functioning in one aspect



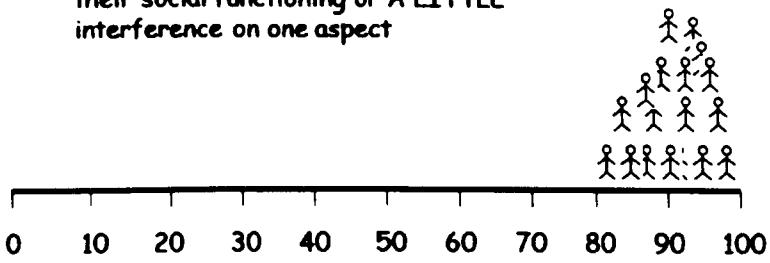
Worst social functioning

Best social functioning

65 and over 6 months after chemotherapy

Average social functioning score = 92

An average score of 92 means that patients either had no interference with their social functioning or A LITTLE interference on one aspect



Worst social functioning

Best social functioning