VOLUME 29 · NUMBER 1 · JANUARY 1 2011

JOURNAL OF CLINICAL ONCOLOGY

ORIGINAL REPORT

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

Kim Cocks, Madeleine T. King, Galina Velikova, Marrissa Martyn St-James, Peter M. Fayers, and Julia M. Brown

From the Clinical Trials Research Unit, University of Leeds; Cancer Research UK Centre, University of Leeds, St James's Institute of Oncology, Leeds; University of Aberdeen, United Kingdom; Quality of Life Office, Psycho-Oncology Cooperative Research Group, University of Sydney; Centre for Health Economics Research and Evaluation, University of Technology, Sydney, Australia; and the Faculty of Medicine, NTNU, Trondheim, Norway.

Submitted January 25, 2010; accepted September 2, 2010; published online ahead of print at www.jco.org on November 22, 2010.

Written on behalf of the Evidence-Based Interpretation Guidelines Collaborative Group.

Supported by Grant No. C7852/A5653 from the Cancer Research UK.

Presented in part in poster format at the annual meeting of the International Society for Quality of Life Research, New Orleans, LA, October 28-31, 2009.

Authors' disclosures of potential conflicts of interest and author contributions are found at the end of this article.

Corresponding author: Kim Cocks, MSc, Clinical Trials Research Unit, University of Leeds, United Kingdom, LS2 9JT; e-mail: kcstats@aol.com.

0732-183X/11/2901-89/$20.00

DOI: 10.1200/JCO.2010.28.0107

ABSTRACT

Purpose
To use published literature to estimate large, medium, and small differences in quality of life (QOL) data from 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 estimate large, medium, and small differences for QLQ-C30 scores. Published mean data were identified from the literature. Differences (contrasts) between groups (eg, between treatment groups, age groups, and performance status groups) were reviewed by 34 experts in QOL measurement and cancer treatment. The experts, blinded to actual QOL results, were asked to predict these differences. A large difference was defined as one representing unequivocal clinical relevance. A medium difference was defined as likely to be clinically relevant but to a lesser extent. A small difference was one believed to be subtle but nevertheless clinically relevant. A trivial difference was used to describe circumstances unlikely to have any clinical relevance. Actual QOL results were combined using meta-analytic techniques to estimate differences corresponding to small, medium, or large effects.

Results
Nine hundred eleven articles were identified, leading to 152 relevant articles (2,217 contrasts) being reviewed by at least two experts. Resulting estimates from the meta-analysis varied depending on the subscale. Thus, the recommended minimum to detect medium differences ranges from 9 (cognitive functioning) to 19 points (role functioning).

Conclusion
Guidelines for the size of effects are provided for the QLQ-C30 subscales. These guidelines can be used for sample size calculations for clinical trials and can also be used to aid interpretation of differences in QLQ-C30 scores.

J Clin Oncol 29:89-96. © 2010 by American Society of Clinical Oncology

INTRODUCTION

The European Organisation for the Research and Treatment of Cancer Quality of Life Questionnaire Core 30 (EORTC QLQ-C30) is one of the most widely used instruments for assessing health-related quality of life (QOL) in patients with cancer. A review[1] of randomized controlled trials (RCTs) highlighted that although studies using the QLQ-C30 were reported to a high standard, clinical interpretation of QOL differences was lacking (62% not addressing clinical significance). There was an over-reliance on statistical significance to determine impact on QOL. Where clinical significance was addressed, it was most common to assume that 10 points[2] was clinically significant (however, fewer than 25% of the RCTs used this method). Reporting of sample size calculations was also lacking. Twelve RCTs specified QOL as a primary end point; however, only seven detailed the sample size calculation. There was no consistent basis for these calculations.

Several authors[3,4] provide guidelines for sample size/interpretation of QOL scores applicable regardless of instrument. There are also a limited number of guidelines specifically for the QLQ-C30.

King[5] estimated effect sizes using clinical anchors from 14 studies and compared with Cohen's.[6] Population-based reference values are available.[7-10] Osoba et al[2] published small/moderate/large changes in scores based on global ratings of change (limited to breast/lung cancer and four QLQ-C30 subscales). The observations on size of change from this study were similar to King.[5]

Despite availability of guidelines, our review showed these are not being utilized.[1] As a consequence, QOL results rarely influence clinical practice.[11-14] The Evidence-Based Interpretation Guidelines project aims to improve the current guidelines for sample size calculation and interpretation. Our innovative methods utilize published study results of QLQ-C30 data, and using meta-analytic techniques combine these with blinded expert opinions as the basis for estimating a clinically relevant change.[15] This robust evidence-based methodology should encourage use of the guidelines at the study design stage and when interpreting scores from the QLQ-C30.

METHODS

Selection of Relevant Articles

Potential QLQ-C30 data sources were identified by searching CINAHL/Medline/Embase/Medline-in-process and Psychinfo databases (post 1998). The EORTC bibliography[16] supplemented the search.

Relevant sources provided mean differences for an informative contrast (comparing two independent groups [a cross-sectional contrast] or the mean change within a group over time [a longitudinal contrast]). We focus on cross-sectional contrasts, of particular importance for the design, sample-size estimation, and interpretation of RCTs. Data sources were reviewed for inclusion by two people.

Expert Panel Reviews

Opinion on whether the effect for contrasts was likely to be trivial, small, medium, or large was sought from a panel of health professionals (ie, experts). Experts were invited if they had experience of treating/caring for patients with cancer and using QLQ-C30. Experts were primarily from EORTC QOL Group

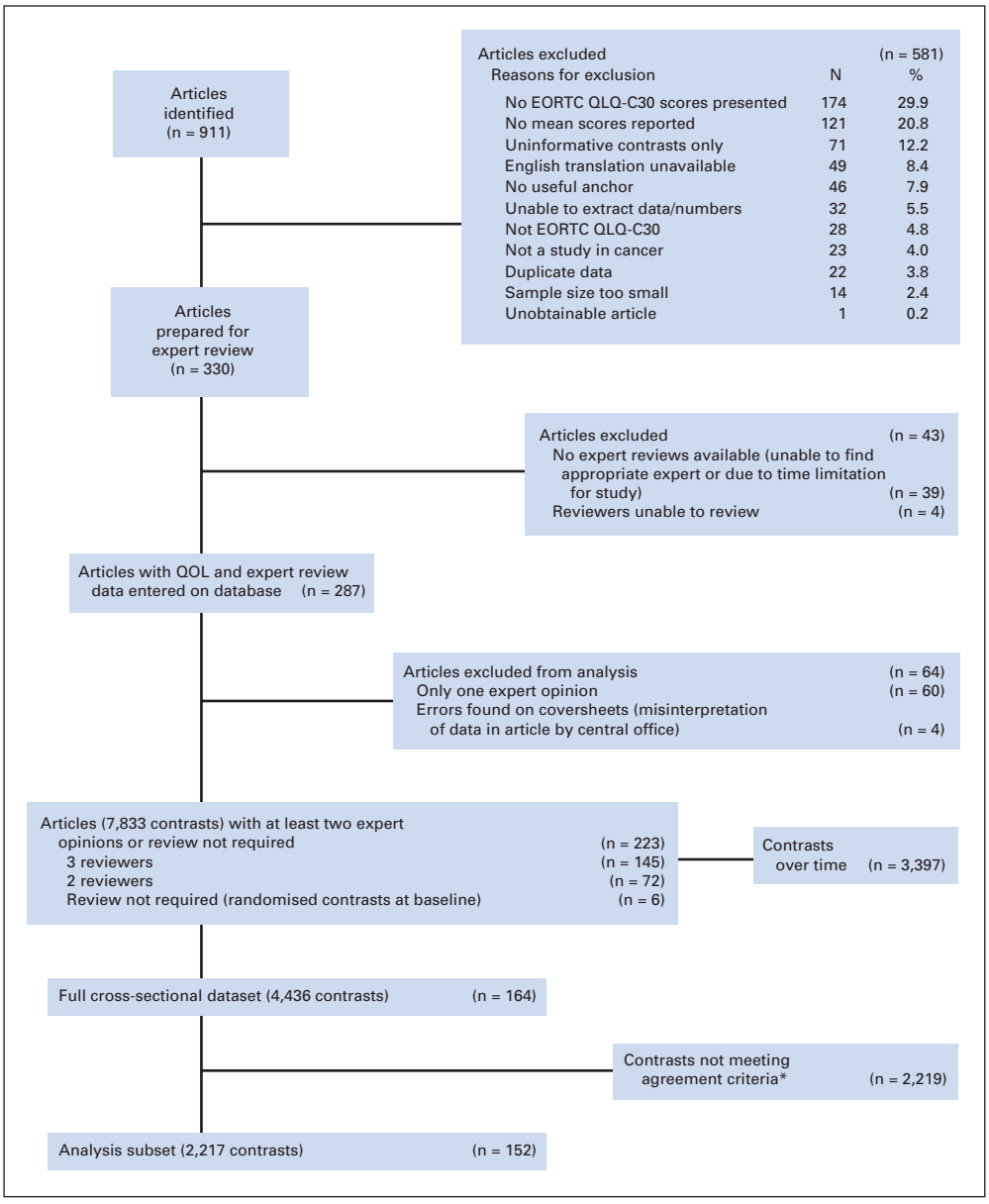

Fig 1. Flow diagram showing flow of articles through the project. (*) Agreement criteria: (1) for trivial contrasts; all experts in agreement. (2) For small, medium, or large contrasts; direction of expert opinions in agreement with the observed direction in the article (eg, experts and article both indicate group A better than group B). EORTC QLQ-C30, European Organisation for the Research and Treatment of Cancer Quality of Life Questionnaire Core 30; QOL, quality of life.

or United Kingdom National Cancer Research Institute Clinical Studies Groups. The panel reviewed articles over a period of 3 years.

We aimed to have experts representing different oncology treatment modalities and sites, and to obtain three opinions per article. Articles were allocated to experts based on the area of expertise required and availability. All sections containing QOL results or interpretation were blacked out before the review. Experts were asked to assess how the QLQ-C30 would behave in that clinical situation. They were asked to judge on the relative size of the effect, as measured by the QLQ-C30. They used four size classes: trivial, small, medium, or large. Large: one representing unequivocal clinical relevance. Medium: likely to be clinically relevant but to a lesser extent. Small: subtle but nevertheless clinically relevant. Trivial: circumstances unlikely to have any clinical relevance or where there was no difference. Fuller definitions available at http://ctro.leeds.uc.uk.

Contrasts were based on clinical anchors, defined as "an independent standard… interpretable and at least moderately correlated with the instrument".[17] All authors reviewed anchors for relevance to QOL before inclusion. If an article contained only randomized contrasts at baseline, these were assigned as trivial since observed differences should only be due to chance. Examples of anchors/contrasts can be found in the Data Supplement.

Reviewers used a scale from −3 to 3; 3 = large, 2 = medium, 1 = small and 0 = trivial. The negative/positive indicated the direction of effect (ie, group A better than group B or vice versa).

Reviewers used percentages to indicate their certainty. If they were confident the observed difference would be in a particular size class they could assign 100%, or they could spread their expectation, with the percentages across size classes summing to 100%. The expert panel instruction manual containing full details is provided in Appendix Figure A1, online only.

Analysis Methods

Observed mean differences. Observed mean differences in reported QOL scores were summarized to show the range found in the literature. The proportion of contrasts meeting the commonly used 10 points criteria for a moderate difference is reported. The subset of randomized treatment comparisons were summarized separately as these may be informative when considering sample size calculations for RCTs.

Expert size classes. For each reviewer, on each contrast, their weighted average was calculated using their percentage certainty as weights; referred to as an individual opinion. The mean of these weighted averages was then calculated for each contrast, referred to as overall opinion for that contrast. The overall opinion was categorized into trivial/small/medium/large; referred to as expert size class. Reviews where one individual opinion was in the opposite direction to another reviewer were sent back to all reviewers for checking. A quality assessment for the meta-analysis considered agreement between reviewers on the same contrast and agreement between the overall opinion and the observed mean difference. Agreement between reviewers was summarized using proportion of contrasts with maximum distance between reviewers of 1, 2, 3, or more categories. Correlation (Spearman's rank) was used to measure association between the overall opinion and observed differences.

Agreement criteria, the full data set, and the analysis subset. The term full data set is used to describe the articles/contrasts with at least one expert review or where expert review was not required (randomized contrasts at baseline). The analysis subset contains the subset of contrasts which were reviewed by more than one expert; only these were used to define the guidelines. The analysis subset was further refined using two distinct agreement criteria. The first criterion applied to trivial contrasts. Only contrasts where all experts agreed were included. This was to reduce the chances of including QOL results with substantial differences in this category. The second criterion applied to contrasts in the other size classes. The subset where overall opinion was in the opposite direction to the article was excluded. Including these would have resulted in artificially reducing estimates for each size class.

Meta-analysis. Meta-analysis methods were used to pool the QOL results using both mean differences and effect sizes as outcome variables. Effect size was calculated as the mean difference divided by the best available estimate of between-person standard deviation. When standard deviation was not reported, it was derived or imputed as the weighted average of available estimates for each subscale separately.[18] 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. Because of the heterogeneity in the contrasts being pooled here (ie, across different clinical anchors) a random effects model was required.[19] This is a standard method for dealing with heterogeneity in a meta-analysis. Contrasts were nested within articles (random effect) to account for possible correlation between contrasts from the same article. The expert size class was the fixed effect in the model. Models were estimated for each subscale separately. Appendix Fig A1 (online only) shows an example forest plot to illustrate the analysis methods further.

Table 1. Characteristics of Reviewed Articles

Characteristic	Full Data Set (n = 164) No.	%	Analysis Data Set (n = 152) No.	%
Cancer site				
Breast	39	23.8	37	24.3
Lung	24	14.6	24	15.8
Head and neck	19	11.6	18	11.8
Hematologic	18	11.0	14	9.2
Colorectal	17	10.4	14	9.2
Multiple	16	9.8	15	9.9
Prostate	14	8.5	14	9.2
GI	7	4.3	6	3.9
Urology/kidney	4	2.4	4	2.6
Testicular	3	1.8	3	2.0
Brain	3	1.8	3	2.0
Research question				
Describe effect of disease and/or treatment on HRQOL	133	81.1	123	80.9
Long-term follow-up of survivors	11	6.7	10	6.6
Develop and/or validate QLQ core module	6	3.7	6	3.9
Psychosocial interventions (eg, nursing, education, counselling programs)	4	2.4	4	2.6
Develop and/or validate QLQ core and disease specific module	3	1.8	3	2.0
Comparison of HRQOL instruments	2	1.2	2	1.3
Cross-cultural validation of the QLQ-C30	2	1.2	2	1.3
Develop and/or validate another HRQOL questionnaire	1	0.6	1	0.7
Cultural issues in HRQOL and health care	1	0.6	1	0.7
Relationship between HRQOL and other variables (eg, age, sex, survival)	1	0.6	0	
Study design				
Cohort/descriptive study	110	67.1	105	69.1
RCT phase not specified	31	18.9	25	16.4
RCT phase III	13	7.9	13	8.6
Multiple studies	7	4.3	7	4.6
Phase II	2	1.2	1	0.7
Phase III crossover	1	0.6	1	0.7
Region				
Europe	114	69.5	103	67.8
United States/Canada	22	13.4	22	14.5
Rest of world	28	17.1	27	17.8
Sample sizes				
Median	172		176	
Range	12-2,640		12-2,640	

Abbreviations: HRQOL, health-related quality of life; QLQ, European Organisation for the Research and Treatment of Cancer Quality of Life Questionnaire; QLQ-30, European Organisation for the Research and Treatment of Cancer Quality of Life Questionnaire Core 30; RCT, randomized controlled trial.

Guidelines for sample size and interpretation. The meta-analysis was 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 9 points for a medium effect then the guidelines would recommend that 7 points be used as a minimum medium effect for sample size calculations/interpretation.

RESULTS

Relevant Sources

Nine hundred eleven articles were identified in the literature review, 581 were excluded (Fig 1); 287 articles were subsequently evaluated by the experts. There were 164 articles in the full data set (55 RCTs), and 152 articles (2,217 individual contrasts) in the analysis subset.

Characteristics of the studies/patients contributing to the analysis are summarized in Table 1. The contrasts were related to a wide range of anchors which were broadly categorized as: treatment/intervention, time of follow-up, disease, patient characteristics, function/symptom, or survival. The number of contrasts ranged from 1 to 14 per article (median, 2).

Expert Panel Reviews

Thirty-four experts reviewed from one to 98 articles each (median, 12). Reviewers were mainly oncologists but the panel also contained nursing, psychosocial, surgical, psychology, and radiotherapy expertise. For the full data set, the proportion of contrasts with reviewers in exact agreement was 38%. A further 39% of contrasts had a maximum distance of one category between reviewers. Only 6% had reviewers three or more categories apart in their opinion. For the analysis data set, the proportion of contrasts with reviewers in exact agreement increased to 49%. Correlation of mean review scores with observed differences was 0.25 for the full data set, improving to 0.62 in the analysis data set.

Observed Mean Differences

Observed mean differences between groups are summarized in Table 2. The full data set, rather than the analysis data set, is used here to show the range of differences found in the literature. The average mean difference ranged from 4 to 11 points across subscales. Physical/role functioning had the widest range (0 to 60 points and 64 points, respectively). Cognitive functioning had the smallest range of mean differences (0 to 37 points); 21% to 41% of contrasts across subscales reached differences of 10 points or more, with highest proportions for fatigue (33%) and role functioning (41%).

Observed differences from the subset of RCTs are also summarized because these may be of particular interest when considering sample size calculations for new RCTs. Mean differences for RCTs were generally smaller than for other study designs (3 to 7 points; Table 2). Unlike the full data set, pain/social functioning had the widest range of mean differences (0 to 30 points) and emotional functioning the narrowest range (0 to 12 points). Four percent to 28% of the contrasts reached differences of 10 points or more. Pain, role, and social functioning had the highest proportion reaching 10 or more points (25% to 28%). Cognitive and emotional functioning rarely showed differences of more than 10 points (4% and 5%, respectively).

Meta-Analysis Results

Figures 2 and 3 show estimates for mean differences/effect sizes, respectively. Trivial, small, and medium estimates with 95% CIs are displayed. Table 3 shows the number of contrasts contributing to the analysis for each estimate. Most subscales show clear trends across size categories, with an increase in estimates from trivial through to medium. Role functioning shows the widest range between the estimates

Table 2. Descriptive Summary of Observed Mean Differences Reported in Articles

	Full Data Set									RCT Treatment Comparisons*								
Subscale	Mean	Median	SD	SE	Max	Min	No.	No. > 10 Points	%	Mean	Median	SD	SE	Max	Min	No.	No. > 10 Points	%
AP	7.87	5.50	7.88	0.49	41.35	0.00	263	68	26	6.43	4.35	6.75	1.04	35.60	0.08	42	11	26
CF	6.48	4.60	6.32	0.36	36.67	0.00	310	66	21	3.47	3.00	3.22	0.51	16.10	0.00	40	2	5
CO	7.41	5.00	7.46	0.48	45.67	0.00	241	64	27	5.11	3.15	4.75	0.77	16.60	0.00	38	6	16
DI	4.44	3.00	4.21	0.28	22.00	0.00	230	20	9	2.99	2.25	2.99	0.48	12.00	0.00	38	1	3
DY	7.07	5.20	5.90	0.39	31.60	0.00	227	56	25	4.46	3.80	4.64	0.80	24.00	0.00	34	3	9
EF	6.34	4.20	6.19	0.33	42.08	0.00	360	77	21	3.29	2.66	2.82	0.41	12.30	0.00	48	2	4
FA	8.81	6.25	8.06	0.46	47.00	0.00	310	103	33	5.94	5.00	4.93	0.68	23.60	0.00	52	7	13
FI	7.09	5.10	6.67	0.46	42.60	0.00	206	51	25	4.04	4.30	2.89	0.60	11.90	0.17	23	1	4
NV	5.15	3.00	6.10	0.38	38.00	0.00	263	40	15	5.52	3.46	6.69	1.02	35.00	0.00	43	8	19
PA	8.42	6.00	8.45	0.49	52.00	0.00	303	86	28	6.24	5.64	5.41	0.82	30.00	0.00	44	11	25
PF	9.32	6.00	9.75	0.52	60.00	0.00	352	102	29	5.10	4.00	4.74	0.67	20.00	0.00	50	8	16
QL	7.22	5.00	7.22	0.36	43.00	0.00	407	102	25	4.29	3.30	4.45	0.57	18.90	0.00	61	8	13
RF	11.19	8.00	10.83	0.58	64.00	0.00	343	142	41	6.95	6.10	5.57	0.81	23.00	0.00	47	13	28
SF	7.82	6.00	7.19	0.37	50.00	0.00	369	98	27	6.27	5.50	5.50	0.79	30.80	0.00	49	13	27
SL	6.88	5.70	5.75	0.36	29.30	0.00	252	58	23	5.31	4.50	3.77	0.59	16.00	0.20	41	5	12
All	7.56	5.00	7.68	0.12	64.00	0.00	4436	1,133	26	5.09	4.00	4.92	0.19	35.60	0.00	650	99	

Abbreviations: RCT, randomized controlled trial; SD, standard deviation; AP, appetite loss; CF, cognitive functioning; CO, constipation; DI, diarrhea; DY, dyspnea; EF, emotional functioning; FA, fatigue; FI, financial difficulties; NV, nausea and vomiting; PA, pain; PF, physical functioning; QL, global quality of life; RF, role functioning; SF, social functioning; SL, insomnia.

*There are 55 RCTs, some of which do not have a treatment comparison for all subscales and some have multiple treatment comparisons within a RCT.

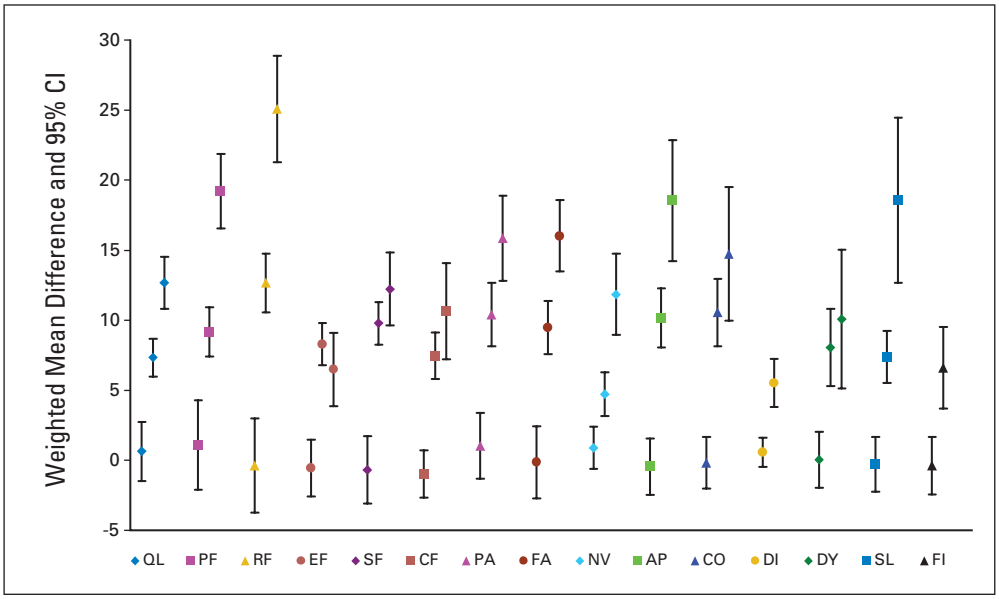

Fig 2. Estimates of mean differences in quality of life for trivial, small, and medium effects from random effects meta-analysis of cross-sectional contrasts.

(mean difference of 0, 13, and 25 points for trivial, small, and medium, respectively), while other subscales, such as global QOL have a smaller range of estimates (mean difference of 1, 7, and 13, respectively). For the emotional, social, cognitive, constipation, and dyspnea subscales there is some degree of overlap between the CIs for the small/medium estimates.

Guidelines for Sample Size Calculations and Interpretation

Guidelines for trivial, small, medium, and large effects are provided in Table 4 for both mean differences/effect sizes. Using the global QOL scale as an example, estimates for trivial, small, and medium mean differences were 1, 7, and 13 points respectively (Fig 2). The

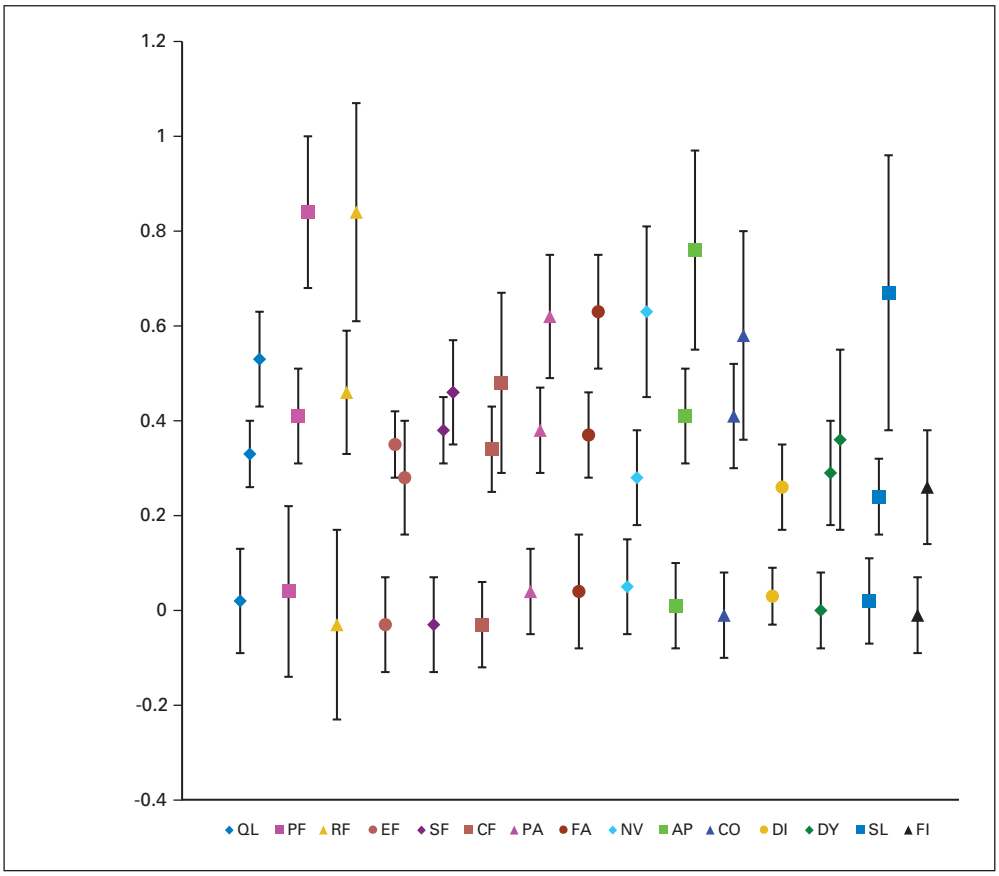

Fig 3. Estimates of effect sizes for trivial, small, and medium effects from random effects meta-analysis of cross-sectional contrasts.

Table 3. Number of Contrasts in Meta-Analysis

Subscale Abbreviation	Subscale	No. of Trivial Contrasts	No. of Small Contrasts	No. of Medium Contrasts
QL	Global quality of life	39	97	54
PF	Physical functioning	32	102	48
RF	Role functioning	40	96	32
EF	Emotional functioning	54	91	33
SF	Social functioning	42	101	36
CF	Cognitive functioning	70	74	19
PA	Pain	53	62	30
FA	Fatigue	38	71	39
NV	Nausea and vomiting	58	56	17
AP	Appetite loss	60	56	14
CO	Constipation	72	44	11
DI	Diarrhea	96	38	1*
DY	Dyspnea	62	38	12
SL	Insomnia	45	57	6
FI	Financial difficulties	54	25	3*

*Results not displayed on Figures 2 and 3 where n < 5.

threshold between size classes is at the midpoint between estimates. The threshold between trivial and small is 4 points (ie, the midpoint between 1 and 7), and between small and medium differences is 10 points (ie, at the midpoint between 7 and 13). In order to use these guidelines to calculate a sample size (assuming it is required to detect the smallest medium difference), this threshold of 10 points should be used in the calculation. To use the guidelines for interpretation, an observed difference of fewer than 4 points, for example, would be interpreted as trivial. The emotional functioning subscale has been omitted due to the medium estimate being lower than the estimate for small effects. Although there was insufficient data to estimate the size of large effects, the upper limit of the 95% CIs around the medium estimates have been used as a guide.

DISCUSSION

Using innovative methodology based on high-quality QOL studies, expert opinions, and meta-analytic techniques, we have produced evidence-based guidelines for trivial to large QLQ-C30 QOL differences (Table 4). Estimates utilize 2,000+ contrasts from published scores and incorporate reviews from 34 cancer/QOL experts. These new estimates highlight previous guidelines may be too simplistic in that they do not distinguish between subscales. We add to the current literature by providing guidelines for each QLQ-C30 subscale. Researchers can now more accurately calculate sample size according to the subscale of primary interest and interpret QOL differences between groups of patients. These guidelines will be more widely applicable than those currently available as they are based on a wide range of cancers and clinical situations. We focus here on cross-sectional contrasts which are informative for sample size calculations and interpretation of between group differences. The longitudinal contrasts will be reported separately as the application of this method to contrasts affected by response shift/attrition warrants separate discussion.

We suggest the threshold between trivial/small would be the smallest estimate on which to base a sample size. Depending on the individual study/interventions larger differences may be of interest and the range of small/medium estimates could be used. It was rare for the experts to expect large differences between groups even when comparing very distinct groups of patients. This lack of large differences has also been observed in studies of patients over time.[2,20] If we retrospectively apply our guidelines to the RCTs, large effects are observed in only 14 (2%). Therefore, researchers designing a study should consider if large effects can reasonably be expected.

Our estimates of medium effects lie in the same range as suggested by Osoba et al[2] for global QOL (10 to 15 points), social (11 to 15 points), and pain and fatigue (13 to 19 points) subscales. Compared with King[21] our results for small are very similar for physical, role, cognitive, nausea, and pain. Our study and King's considered group

Table 4. Guidelines for Size of Cross-Sectional Differences (from meta-analysis)

Lower Estimate of Medium Differences (points)*	Subscale	Mean Difference				Effect Size†			
		Trivial	Small	Medium	Large	Trivial	Small	Medium	Large
< 10	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	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	RF	0-6	6-19	19-29	> 29	0-0.2	0.2-0.7	0.7-1.1	> 1.1

Abbreviations: DI, diarrhea; NV, nausea and vomiting; CF, cognitive functioning; DY, dyspnea; FI, financial difficulties; QL, global quality of life; SF, social functioning; SL, insomnia; FA, fatigue; CO, constipation; PA, pain; PF, physical functioning; AP, appetite loss; RF, role functioning.

*Subscales have been ordered in this Table according to the size of the medium differences.

†Effect size refers to the standardized mean difference (ie, mean difference divided by the best available estimate of between-person standard deviation).

differences using published data whereas Osoba et al used individual patients' ratings of change over time to produce guidelines. Despite these differences, there is substantial overlap in the resulting guidelines from the three studies, which is reassuring.

We used experts to estimate impact on patients' QOL. Experts have previously been shown to underestimate symptom severity, but in the same study patients and doctors showed similar conclusions with respect to between treatment differences.[22] We believe the use of experts is justified 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. We conducted a pilot study to research feasibility of incorporating patient opinions on published data in a similar way. Full results are not reported here. We found that 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. Therefore, a larger panel of patients would be required compared to a panel of experts (who have a broader range of clinical experiences) in order to judge a wide range of clinical settings. However, we also found patients' views were generally in the same size class or occasionally a larger size class than experts. It therefore seems reasonable to assume expert opinions are adequate and appropriate here.

Our analysis data set contained around half the identified contrasts. The similar Functional Assessment of Cancer Therapy General Scale (FACT-G) project[15] highlighted contrasts with agreement between the experts as important for validity of results, however, it was unclear which study/contrast characteristics led to good agreement. Therefore, for this study, we did not set stringent criteria for inclusion but later applied criterion to exclude contrasts where experts disagreed markedly with each other or with the observed results. As a result of our inclusivity, the correlation between experts and observed differences was low for the full data set. Further work is being carried out to identify the study characteristics leading to poor agreement. Post hoc exclusion is a weakness in our study, future studies would benefit from excluding these contrasts up front.

Our results showed some overlap between the small/medium estimates for emotional, social, cognitive, constipation, and dyspnea subscales. Emotional functioning could not be included in the guidelines for sample size calculations. This may be an indication that the 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. It is likely that the larger changes in these subscales would arise from psychosocial-related anchors/interventions which are not common in the literature. When planning a study, it is likely that these subscales are appropriate as primary end points only in psychosocial interventions, as they are unlikely to be changed systematically in other situations.

Our guidelines are based on published data. We observed, in common with others, a substantial amount of variation between the mean differences which are pooled by the experts into the same size class. This highlights the need for careful consideration at the design stage of a study of the factors that may affect the QOL differences (eg, timing of QOL assessments), nature of interventions, and subscales of primary interest. These guidelines can be used for sample size calculations for clinical trials and can also be used to aid interpretation of differences in QLQ-C30 scores.

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The author(s) indicated no potential conflicts of interest.

AUTHOR CONTRIBUTIONS

Conception and design: Kim Cocks, Madeleine T. King, Galina Velikova, Peter M. Fayers, Julia M. Brown
Provision of study materials or patients: Galina Velikova
Collection and assembly of data: Kim Cocks, Galina Velikova, Marrissa Martyn St-James
Data analysis and interpretation: Kim Cocks, Madeleine T. King, Galina Velikova, Peter M. Fayers, Julia M. Brown
Manuscript writing: All authors
Final approval of manuscript: All authors

REFERENCES

1. Cocks K, King MT, Velikova G, et al: 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 44:1793-1798, 2008
2. Osoba D, Rodrigues G, Myles J, et al: Interpreting the significance of changes in health-related quality-of-life scores. J Clin Oncol 16:139-144, 1998
3. Wyrwich KW: Minimal important difference thresholds and the standard error of measurement: Is there a connection? J Biopharm Stat 14:97-110, 2004
4. Norman GR, Sloan JA, Wyrwich KW: Interpretation of changes in health-related quality of life: The remarkable universality of half a standard deviation. Med Care 41:582-592, 2003
5. King MT: Cohen confirmed? Empirical effect sizes for the QLQ-C30. Qual Life Res 10:278, 2001
6. Cohen J: Statistical power analysis for the behavioral sciences (rev ed). Hillsdale, NJ, Lawrence Erlbaum Associates Inc, 1977
7. Hjermstad MJ, Fayers PM, Bjordal K, et al: 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 16:1188-1196, 1998
8. Klee M, Groenvold M, Machin D: Quality of life of Danish women: Population-based norms of the EORTC QLQ-C30. Qual Life Res 6:27-34, 1997
9. Michelson H, Bolund C, Nilsson B, et al: Health-related quality of life measured by the EORTC QLQ-C30–reference values from a large sample of Swedish population. Acta Oncol 39:477-484, 2000
10. Schwarz R, Hinz A: Reference data for the quality of life questionnaire EORTC QLQ-C30 in the general German population. Eur J Cancer 37:1345-1351, 2001
11. Blazeby JM, Avery K, Sprangers M, et al: Health-related quality of life measurement in randomized clinical trials in surgical oncology. J Clin Oncol 24:3178-3186, 2006
12. Efficace F, Bottomley A, Osoba D, 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? J Clin Oncol 21:3502-3511, 2003
13. Goodwin PJ, Black JT, Bordeleau LJ, et al: Health-related quality-of-life measurement in randomized clinical trials in breast cancer: Taking stock. J Natl Cancer Inst 95:263-281, 2003
14. Guyatt GH, Schunemann HJ: How can quality of life researchers make their work more useful to health workers and their patients? Qual Life Res 16:1097-1105, 2007
15. King MT, Stockler MS, Cella D, et al: Meta-analysis provides evidence-based effect sizes for a cancer-specific quality of life questionnaire, the FACT-G. J Clin Epidemiol 63:270-281, 2010

16. European Organisation for the Research and Treatment of Cancer: Bibliography. http://groups.eortc.be/qol/documentation_bibliography.htm

17. Guyatt GH, Osoba D, Wu AW, et al: Methods to explain the clinical significance of health status measures. Mayo Clin Proc 77:371-383, 2002

18. Follmann D, Elliott P, Suh I, et al: Variance imputation for overviews of clinical trials with continuous response. J Clin Epidemiol 45:769-773, 1992

19. Lipsey MW, Wilson DB: Practical Meta-Analysis. Thousand Oaks, CA, Sage Publications Inc, 2001

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

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

22. Stephens RJ, Hopwood P, Girling DJ, et al: Randomized trials with quality of life endpoints: Are doctors' ratings of patients' physical symptoms interchangeable with patients' self-ratings? Qual Life Res 6:225-236, 1997

