Patient Related Outcome Measures

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

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

This article was published in the following Dove Press journal:
Patient Related Outcome Measures
22 September 2010
Number of times this article has been viewed

Madeleine T King[1]
David Cella[2]
David Osoba[3]
Martin Stockler[4]
David Eton[5]
Joanna Thompson[6]
Amy Eisenstein[7]

[1]Psycho-oncology Co-operative Research Group School of Psychology, University of Sydney, New South Wales, Australia; [2]Department of Medical Social Sciences, Northwestern University, Illinois, USA; [3]QOL Consulting, Vancouver, British Columbia, Canada; [4]NHMRC Clinical Trials Centre, University of Sydney, New South Wales, Australia; [5]Mayo Clinic, Rochester, Minnesota, USA; [6]Centre for Health Economics Research and Evaluation, University of Technology, Sydney, New South Wales, Australia; [7]Center on Outcomes Research and Education (CORE), Evanston Northwestern Healthcare (ENH), Evanston, Illinois, USA

Abstract: Our aim was to develop evidence-based interpretation guidelines for the Functional Assessment of Cancer Therapy-General (FACT-G), a cancer-specific health-related quality of life (HRQOL) instrument, from a range of clinically relevant anchors, incorporating expert judgment about clinical significance. Three clinicians with many years' experience managing cancer patients and using HRQOL outcomes in clinical research reviewed 71 papers. Blinded to the FACT-G results, they considered the clinical anchors associated with each FACT-G mean difference, predicted which dimensions of HRQOL would be affected, and whether the effects would be trivial, small, moderate, or large. These size classes were defined in terms of clinical relevance. The experts' judgments were then linked with FACT-G mean differences, and inverse-variance weighted mean differences were calculated for each size class. Small, medium, and large differences (95% confidence interval) from 1,118 cross-sectional comparisons were as follows: physical well-being 1.9 (0.6–3.2), 4.1 (2.7–5.5), 8.7 (5.2–12); functional well-being 2.0 (0.5–3.5), 3.8 (2.0–5.5), 8.8 (4.3–13); emotional well-being 1.0 (0.1–2.6), 1.9 (0.3–3.5), no large differences; social well-being 0.7 (–0.7 to 2.1), 0.8 (–2.9 to 4.5), no large differences. Results from 436 longitudinal comparisons tended to be smaller than the corresponding cross-sectional results. These results augment other interpretation guidelines for FACT-G with information on sample size, power calculations, and interpretation of cancer clinical trials that use FACT-G.

Keywords: health-related quality of life, patient-reported outcomes

Introduction

Health-related quality of life (HRQOL) questionnaires have the potential to play a key role in bringing the patient's voice to evidence-based health care. But to realize this potential, we need to interpret the relevance of HRQOL outcomes to decisions about treatment. Such decisions are made at both the individual level, when a patient (or patient's clinician, acting as patient's agent) chooses among treatment options, and the group level, when clinical research is conducted to test the effectiveness of new treatments relative to current routine treatment.[1,2] In this article, we focus on the latter. For example, if a clinical trial shows that a new treatment improves the mean HRQOL by 10 points on a 100-point scale at no extra cost and with no adverse effects relative to the current best treatment, which improves mean HRQOL by only 5 points, we need to know whether this difference is big enough to change policy and practice. Investigators planning new studies need this type of knowledge to calculate sample sizes, and end-users need to understand the implications of study results for future clinical practice and policy.

Correspondence: Madeleine T King
Psycho-oncology Co-operative Research Group (PoCoG), School of Psychology, Brennan MacCallum Bldg (A18), University of Sydney,
NSW 2006, Australia
Tel +61 2 9036 6114
Fax +61 2 9036 5292
Email madeleine.king@sydney.edu.au

submit your manuscript | www.dovepress.com
Dovepress
DOI: 10.2147/PROM.S10621

Despite increasing acceptance and use of HRQOL questionnaires as valid and informative measures in clinical research, interpreting the clinical relevance of outcomes measured on their scales remains challenging.[3] HRQOL is multifaceted and subjective, and there are a large number and wide range of measurement scales, each of which has a different and somewhat arbitrary scale. Few people have the requisite understanding of psychometrics or the hands-on experience to confidently interpret specific HRQOL scales. Familiarity with a measurement scale, whether it measures physical phenomena such as temperature or perception-based phenomena such as HRQOL, is the key to intuitive understanding of results recorded on that scale. Such familiarity develops with experience of how the numbers from such scales correspond to events in our lives and the significance of those to our actions and decisions. The multitude of HRQOL instruments compounds the problem of developing familiarity with specific HRQOL scales. Some HRQOL questionnaires are now very widely used. The 2 most commonly used cancer-specific instruments are the European Organisation for Research and Treatment of Cancer's Quality of Life Questionnaire (QLQ)-C30[4] and the functional assessment of cancer therapy-general (FACT-G).[5] Collective experience with these 2 instruments has amassed a rich evidence base for developing interpretation guidelines.

Clinicians who have specialized in HRQOL research and used particular HRQOL measures over many years in both research and clinical contexts are ideally positioned to develop intuitive interpretations of HRQOL results.[2] Over many years of managing patients, administering HRQOL questionnaires, analyzing and scrutinizing the results, and linking HRQOL scores with the condition of their patients, they understand the variation among patients in their levels of HRQOL, how they react to their disease and treatment, and how their HRQOL scores change over time. They also understand how to interpret the mean scores from groups of such individuals.

Some of the most useful evidence for developing interpretations that are meaningful to clinicians comes from studies that report the HRQOL of patients grouped by established clinical criteria, sometimes called 'clinical anchors'[6] or 'known groups'.[4] This type of data is often collected during the validation of an instrument, with patterns of HRQOL scores across clinical groups providing evidence of clinical validity. Quantitatively, these patterns can be used to develop interpretation guidelines about the relative size and significance of mean differences on HRQOL scales. Similarly, the patterns in longitudinal data collected during conventional treatments with well known and understood clear clinical effects help us understand the relative size and significance of mean changes in HRQOL. This method has been applied to the QLQ-C30, a cancer-specific HRQOL questionnaire.[7] In this article, we further develop this method and apply it to another cancer-specific HRQOL questionnaire, the FACT-G.[5]

The FACT-G forms the central core of a suite of questionnaires, referred to as the Functional Assessment of Chronic Illness Therapy (FACIT). Additional questions focus on specific diagnoses such as lung cancer (FACT-L) or breast cancer (FACT-B), and on symptoms such as fatigue (FACT-F) and anemia (FACT-An). These questionnaires are used increasingly in clinical research, adding to the body of evidence about the FACIT suite. In this article, we focus on the FACT-G, which is summarized as a total score and 4 subscales: physical well-being (PWB), social or family well-being (SWB), emotional well-being (EWB), and functional well-being (FWB).

We have previously described a new approach to synthesizing evidence about HRQOL measures with information on the interpretation of effect sizes derived from the FACT-G.[8] In that paper, we focused on effect sizes, comparing our results with Cohen's guidelines for small, medium, and large effect sizes,[9] and the proposition that a 0.5 SD is the minimum significant difference.[10] The current article presents analogous results for the raw scores from the 5 scales of the FACT-G.

Methods

A detailed description of the methods in relation to effect sizes is given elsewhere.[8] Condensed extracts included here allow readers to assess the meta-analysis of the 5 FACT-G scale scores reported in this article.

Data sources

Papers that reported on the FACT-G were identified by searching relevant online databases. Unpublished information was identified through the FACIT Projects Register. Papers were included if they reported the mean difference at least between 2 independent groups (a cross-sectional contrast) or the mean change within at least 1 group over time (a longitudinal contrast). Results that represented duplicate publication were excluded, as were results from a total sample of less than 10 patients (potentially unreliable) or repeated measures from a sample with greater than 20% attrition (potentially biased).

Expert judgment

The clinical relevance of included mean differences was judged by 3 of us (DC, DO, MS). All identifying information

on papers was obscured, as were any results and conclusions about the FACT-G scores. Each expert predicted the degree of difference in HRQOL for each mean difference between groups or within a group over time. Judgments were constrained to 8 options: much better (3), moderately better (2), a little better (1), much the same (0), a little worse (−1), moderately worse (−2), much worse (−3), and "don't know". After the initial round of judgments, contrasts for which any 2 judges' scores differed by 2 or more categories were reconsidered. The 3 experts worked independently in both the initial and the consensus phases. Weighted kappa was calculated as a measure of concordance of the final judgment scores, and we interpreted kappa values after Landis and Koch.[11] The average of all 3 experts' judgments, rounded to the nearest integer, was used to determine the final size class for each mean difference.

Definitions of size categories

Prior to the panel judging any papers, we defined 4 size groups explicitly in terms of their clinical relevance, where a clinically relevant difference was one that implied a difference in prognosis or clinical management. When circumstances had obvious and unequivocal clinical relevance (eg, patients with asymptomatic, early stage disease vs those with end-stage disease, or the change induced by a treatment well known to markedly improve the health state of most patients treated), group-level HRQOL was expected to be much better or worse (large effect). When circumstances were likely to have clinical relevance but to a lesser extent (eg, for patients with metastatic disease, the contrast of those who were responding to treatment compared with those who were not responding, or a treatment that was known to be effective for half the patients treated), group-level HRQOL was expected to be moderately better or worse (moderate effect). When effects were expected to be subtle but nevertheless clinically relevant (eg, the contrast of patients with regionally advanced cancer vs those with newly diagnosed metastatic disease, or a treatment that was known to improve the health state of only a small proportion of patients treated), group-level HRQOL was expected to be a little better or worse (small effect). When circumstances were unlikely to have any clinical relevance, group-level HRQOL was not expected to be any better or worse (trivial effect).

Data extraction

The following information was extracted from each paper: sample sizes and attrition; standard deviations (required for the inverse-variance weighting factor)[12] or other information from which standard deviations could be derived;[13] the clinical classifications and circumstances of patients (including anchors for mean differences in HRQOL); and mean scores of the PWB, FWB, EWB, and SWB scales. The total score was calculated as the sum of the PWB, FWB, EWB, and SWB scales. Each mean difference was then linked with the corresponding average expert judgment score to allocate a size class for the meta-analysis.

Meta-analysis

Inverse-variance weighted mean differences (IWMDs)[12,14] were calculated for each of 4 size classes by grouping corresponding negative and positive size classes. Thus, all contrasts with a −1 average judgment score were grouped with those with a +1 score. To maintain the correct relationship between the sign of the reported HRQOL differences and the sign of the experts' judgment score, the signs of the HRQOL differences with −1 average judgment scores were reversed prior to grouping them with the +1 contrasts. Results for the medium (−2/+2) and large (−3/+3) contrasts were treated analogously. Cross-sectional contrasts were analyzed separately from longitudinal contrasts.

Sensitivity analysis

We assessed the robustness of results based on all mean differences to discordance between experts by considering only those where at least 2 experts were perfectly concordant, and at most only 1 was discordant by 1 point.

Results

Of 210 papers that satisfied the search criteria, 71 were suitable for inclusion. The full citation details and a summary of the characteristics of these papers can be found in the companion paper.[8] The purpose of the most common study was to describe the effect of disease and treatment on HRQOL (44%), followed by developing and validating a FACIT instrument (21%) and phase 2 clinical trial (7%). A wide range of clinical anchors were reported (Table 1). The most common anchor for differences in HRQOL between groups was routine treatment; 25% of the 71 studies reported mean FACT-G scores by this anchor. The second most commonly used cross-sectional anchor was extent of disease (24%), and the third was performance status (24%) (usually Eastern Cooperative Oncology Group, assessed by the clinician). The most common anchor for change in HRQOL over time was time since starting treatment; 30% of the 71 studies reported mean FACT-G scores by this anchor. The second most commonly used longitudinal anchor was change in

Table 1 Clinical anchors reported in the 71 included studies, and the number of papers that reported mean FACT-G scores by these anchors

Clinical anchor	Cross-sectional	Longitudinal
Treatment group, not randomized	18	–
Disease status/extent	17	–
Performance status	14	8
Time since treatment started	2	21
Fatigue	2	1
Response to therapy	1	1
Other psychological measure	1	1
Chemotherapy	–	2
Global rating of change	–	1
Hemoglobin level	–	1
Survival	–	1
Other HRQOL measure	–	1
Patient location (eg, inpatient, outpatient)	5	–
Gender	3	–
Age	2	–
Diagnostic category	1	–
Exercise	1	–
Ethnic groups	1	–
Language	1	–
Taste changes	1	–
Spirituality	1	–
Time since diagnosis	1	–

Abbreviation: FACT-G, Functional Assessment of Cancer Therapy-General.

performance status (11%). Of the other 22 anchors, most were used in only 1 or 2 papers.

The 71 papers and 22 anchors yielded 1,562 mean differences. For 8 of these, none of the 3 experts were able to make a prediction. In the remainder, the experts differed by 2 or more points for only 17% (261) of their initial judgments. The consensus process reduced this to 6% (95 contrasts). A detailed consideration of expert concordance is given in the companion paper.[8]

Table 2 shows the results of meta-analysis of the raw scale scores from 1,118 cross-sectional contrasts, 436 longitudinal contrasts, and the number of mean differences in each size class. Table 3 shows analogous results from 617 (55% of 1,118) cross-sectional contrasts and 216 (50% of 436) longitudinal contrasts, in which at least 2 experts were perfectly concordant and at least only 1 was discordant by 1 point. The results in Tables 2 and 3 are very similar, indicating that our results are robust to the degree of agreement between experts. Therefore, we focus on Table 2 hereafter. While we cite point estimates here for simplicity, we urge readers to note the range of possible values for each size class and domain suggested by the confidence intervals (CIs) reported in the tables. Across all domains, for both cross-sectional and longitudinal contrasts, IWMDs considered "trivial" by the experts were very small, and their CIs contained zero.

For the other size classes, we consider the cross-sectional results first. The PWB and FWB scales were similar, with IWMDs of 2, 4, and 9 for small, medium, and large effects, respectively. Results for the EWB scale were about half that size, with IWMDs of 1 and 2 for small and medium effects, respectively. Results for the SWB scale were smaller again, there was no gradient from small to medium effects, and all CIs contained zero. For the total FACT-G score, small and medium IWMDs were 6 and 11 points, respectively. Only 2 contrasts were predicted to yield a large effect for the total score, and none was predicted to yield large effects for the EWB or SWB domain.

We now consider the longitudinal results in Table 2. The IWMDs for the PWB and FWB domains were a little less than half the size of corresponding cross-sectional values, with IWMD for small effects being somewhat less than 1, and medium effects having IWMDs of 1.5. Small effects for the EWB scale had an IWMD of about 1. Very little evidence was available to estimate medium effects for the EWB and SWB domains, and there was virtually no evidence available for large effects, with only 1 predicted in the PWB domain. The expected gradient across size classes was most pronounced in the PWB and FWB domains and not apparent at all in the SWB domain. For the SWB domain, all but one of the contrasts was predicted to yield trivial or small effects, and for the small effects, IWMDs were very small and their CIs contained zero.

Discussion

This study (which focuses on FACT-G raw scores) and its companion (which focuses on effect sizes)[8] provide the first formal meta-analysis of anchor-based evidence for a HRQOL instrument, the FACT-G. This evidence covers a wide range of clinically meaningful anchors, and is judged by 3 clinicians with many years of experience managing individual cancer patients and using HRQOL outcomes in cancer clinical trials. In summarizing our results for the FACT-G raw scores, we heed the advice of Guyatt et al[2] to avoid misleading oversimplifications and overly complex presentations. We believe that interpretation guidelines for HRQOL scales require some flexibility to accommodate different patient groups and clinical circumstances, so we summarize our results for each size class and domain as likely ranges. Thus, for the PWB and FWB scales, cross-sectional anchors suggest that a small effect is likely to be in the vicinity of 1–3 points, a medium effect in the vicinity of 3–5 points, and a

Table 2 Results of meta-analysis of the FACT-G raw scale scores

	Physical well-being		Functional well-being		Emotional well-being		Social well-being		Total FACT-G	
Number of contrasts	**X**	**L**	**X**	**L**	**X**	**L**	**X**	**L**	**X**	**L**
Trivial	40	22	46	26	66	22	112	43	33	14
Small	97	32	97	36	122	61	95	42	77	37
Medium	85	33	81	30	44	9	15	1	83	27
Large	13	1	10	0	0	0	0	0	2	0
Cross-sectional (X)	**IWMD**	**95% CI**	**IWMD**	**95% CI**	**IWMD**	**95% CI**	**IWMD**	**95% CI**	**IWMD**	**95% CI**
Trivial	–0.09	–2.2, 2.0	0.3	–2.0, 2.7	–0.16	–1.5, 1.3	0.1	–1.1, 1.4	–1	–8, 6
Small	1.9	0.6, 3.2	2.0	0.5, 3.5	1.0	0.1, 2.6	0.7	–0.7, 2.1	6	2, 11
Medium	4.1	2.7, 5.5	3.8	2.0, 5.5	1.9	0.3, 3.5	0.8	–2.9, 4.5	11	7, 15
Large	8.7	5.2, 12	8.8	4.3, 13	–	–	–	–	22	–4, 48
Longitudinal (L)	**IWMD**	**95% CI**	**IWMD**	**95% CI**	**IWMD**	**95% CI**	**IWMD**	**95% CI**	**IWMD**	**95% CI**
Trivial	0.5	–1.1, 2.0	–0.04	–1.8, 1.7	0.3	–1.1, 1.6	0.02	–1.1, 1.1	0.4	–5.5, 6.4
Small	0.8	–0.4, 2.1	0.7	–0.8, 2.2	1.1	0.3, 1.9	0.17	–0.9, 1.2	2.4	–1.4, 6.1
Medium	1.5	0.1, 2.9	1.5	–0.2, 3.2	0.6	–1.7, 2.8	0.06	–	3.3	–1.3, 7.8
Large	8.2	–	–	–	–	–	–	–	–	–

Note: The scores are from all cross-sectional (X, n = 1,118) informative contrasts and longitudinal (L, n = 436) informative contrasts: number of contrasts and IWMD, with 95% CIs for the 5 FACT-G scales and the 4 size classes.
Abbreviations: FACT-G, Functional Assessment of Cancer Therapy-General; IWMDs, inverse-variance weighted mean differences; CIs, confidence intervals.

large effect in the vicinity of 6–11 points. For these scales, longitudinal anchors yielded smaller estimates, with a moderate effect from longitudinal data being about the size of a small effect from cross-sectional data (1–3 points). For the EWB scale, both cross-sectional and longitudinal anchors suggest that a small effect is likely to be in the vicinity of 1–2 points. Large effects are unlikely to be observed for either the EWB or the SWB, and even small effects may be unlikely for the SWB scale. (CIs for the latter contained zero for the 95 cross-sectional and 42 longitudinal mean differences, expected to yield small effects by the experts). For the total FACT-G score, cross-sectional anchors suggest that a small effect is likely to be in the vicinity of 4–9 points and a medium effect in the vicinity of 9–14 points, with

Table 3 Sensitivity analysis results of meta-analysis of the FACT-G scale scores

	Physical well-being		Functional well-being		Emotional well-being		Social well-being		Total FACT-G	
Number of contrasts	**X**	**L**	**X**	**L**	**X**	**L**	**X**	**L**	**X**	**L**
Trivial	28	14	32	16	51	15	88	39	24	10
Small	56	16	61	20	46	9	37	21	49	18
Medium	41	14	37	13	6	2	5	1	31	7
Large	13	1	10	0	0	0	0	0	2	0
Cross-sectional (X)	**IWMD**	**95% CI**	**IWMD**	**95% CI**	**IWMD**	**95% CI**	**IWMD**	**95% CI**	**IWMD**	**95% CI**
Trivial	0.04	–2.4, 2.5	0.7	–2.1, 3.5	–0.06	–1.5, 1.3	0.2	–1.2, 1.7	–1	–9, 8
Small	2.0	0.3, 3.7	2.2	0.3, 4.1	1.1	–0.4, 2.6	0.9	–1.2, 3.0	7	1, 12
Medium	4.5	2.5, 6.5	4.6	2.1, 7.1	1.8	–2.7, 6.3	0.9	–5.0, 6.7	15	8, 22
Large	8.7	5.2, 12	8.8	4.3, 13	–	–	–	–	22	–4, 48
Longitudinal (L)	**IWMD**	**95% CI**	**IWMD**	**95% CI**	**IWMD**	**95% CI**	**IWMD**	**95% CI**	**IWMD**	**95% CI**
Trivial	0.02	–1.9, 2.0	–0.2	–2.2, 1.9	0.2	–1.2, 1.7	–0.03	–1.1, 1.1	–0.6	–7.5, 6.2
Small	0.9	–0.7, 2.5	0.8	–1.4, 3.0	1.7	–0.7, 4.1	0.4	–1.3, 2.0	2.0	–3.7, 7.7
Medium	2.4	0.3, 4.5	1.2	–1.3, 3.7	2.1	–9.2, 13.5	0.06	–8.5, 8.6	1.1	–7.7, 9.8
Large	8.2	–7.5, 24	–	–	–	–	–	–	–	–

Note: The scores are from the subset of cross-sectional (X, n = 617/1118, 55%) contrasts and longitudinal (L, n = 216/436, 50%) contrasts in which at least 2 experts were perfectly concordant and up to 1 was discordant by at most 1 point: number of contrasts and IWMDs, with 95% CIs for the 5 FACT-G scales and the 4 size classes.
Abbreviations: FACT-G, Functional Assessment of Cancer Therapy-General; IWMDs, inverse-variance weighted mean differences; CIs, confidence intervals.

large effects unlikely. The small effect estimate is similar to the general guideline estimate for the FACT-G minimally important difference (MID), which is 4–7 points.[15] For the total score, results from longitudinal anchors were smaller than corresponding cross-sectional results and were rather unconvincing as interpretation guidelines as their CI contained zero despite reasonable sample sizes.

It is worth noting that for a given size class, the longitudinal mean differences were smaller than cross-sectional ones for the PWB and FWB scales (and consequently the total score). This pattern has been observed previously for the FACT-An and FACT-F scales.[16] The responsiveness of the FACT-G scales has been demonstrated in many papers,[17–26] so we discount lack of responsiveness as an explanation. Several other factors provide more plausible explanations. Response shift, response sets, and other factors related to adjustment may diminish the true size of self-reported change.[27] Another possibility is that our experts overestimated the magnitude of a health domain change likely to be observed with treatment, change of disease course, or other longitudinal anchor. Yet another is that our clinical experts may not have understood the longitudinal anchors as well as they did the cross-sectional anchors; lower rates of concordance among the experts for the longitudinal contrasts than for the cross-sectional contrasts lend some support to this hypothesis. It is also possible that the longitudinal HRQOL assessments may not have occurred at the best time to capture the effects anticipated by the experts. This is a common problem in longitudinal research since HRQOL is commonly assessed at clinic visits, which is convenient and maximizes response rates but does not necessarily capture the peaks and troughs in HRQOL trajectories.[28] Finally, there may have been a minimizing bias introduced by sample attrition; patients who were most likely to deteriorate by the largest amounts were also most likely to be lost to follow-up. In this study, we only included within-group contrasts with less than 20% attrition to minimize this problem. All of these factors made the experts' task of predicting change more difficult than that of separating groups. Some or all of these factors may be working in concert, underlining the challenges in conducting and interpreting longitudinal HRQOL research.

The other finding of interest was that the IWMDs for the social and emotional well-being scales were smaller than those for the physical and functional domains, a pattern evident also for the QLQ-C30.[7] Three general factors may explain these observations. First, the clinical classifications and circumstances prevalent in cancer research may relate more to physical and functional aspects of HRQOL than to psychosocial domains. Second, following the first, our clinical experts may not have been as accurate in their predictions for psychosocial domains as they were for the physical and functional domains. Lower rates of concordance among the experts in the EWB and SWB domains relative to the PWB and FWB domains lend some support to this hypothesis. Third, scales such as the FACT-G's EWB and SWB and the QLQ-C30's emotional, social and cognitive functioning scales may not be as sensitive to real differences in psychosocial aspects of HRQOL as scales such as the FACT-G's PWB and FWB and the QLQ-C30's physical and role functioning scales are sensitive to real differences in physical and functional aspects of HRQOL. We believe the first 2 reasons are more likely than the third since several studies have shown change in emotional well-being using the EWB scale.[5,29–32]

Whatever the reasons for the observation above and if our results generalize to other HRQOL instruments, then the following implications may hold for choice of HRQOL outcomes in cancer trials. HRQOL domains with physical and functional focus may generally yield larger mean differences and hence provide more powerful measures of outcome than do psychosocially focused domains, where at best small effects may be expected. Scales based on social or family well-being may be suitable primary outcomes only for studies of psychosocial interventions targeted specifically at social and family issues in which pilot studies or phase 2 trials have demonstrated an effect for these outcomes.

There was considerable variation in empirical estimates of mean differences within each size category and for each FACT-G scale. There are 2 obvious contributing factors: sampling variation and a degree of mismatch between our experts' expectations and the actual patterns in the HRQOL data. Our 3 experts collectively had a wealth of clinical experience with cancer patients and with HRQOL assessment, so their judgments should be as good as any available. The influence of sampling variation on the outcomes of individual studies cannot be discounted since it is well documented that individuals vary markedly in HRQOL levels at a particular time and in their trajectories of HRQOL over time. The degree of variation of component estimates within size classes in this meta-analysis highlights the limitations of individual studies for deriving general interpretation guidelines.

Other authors have produced evidence across clinical anchors and studies to develop interpretation guidelines.[7,10,15,16,25] Our method advances this type of research in 2 important ways. First, we used formal methods of meta-analysis to produce weighted average mean differences

for each size class. Second, clinical meaningfulness was judged by 3 clinicians with many years' experience managing individual cancer patients and using HRQOL outcomes in cancer clinical trials. Our experts were blinded to the FACT-G scores because we wanted them to place a value on the significance of differences (as determined by the clinical characteristics and circumstances of the patients) rather than to describe the magnitude of differences. Further, our definitions of size classes explicitly address the relevance of HRQOL results to clinical decision making, thereby providing a direct link to Jaeschke et al's widely cited definition of the minimum clinically important difference,[33] more recently modified by Norman et al[10] and Schünemann et al.[34] Rather than focusing on the MID, we accommodate the possibility that in some circumstances, the MID may be of a moderate absolute size while in others it may be relatively small.

The results presented in this article augment other interpretation guidelines for the FACT-G, [15,16,25] adding a substantial evidence base not previously considered for this purpose. We have thereby provided a comprehensive synthesis of anchor-based evidence for the 5 FACT-G scales, incorporating the collective understanding of 3 clinicians with many years' experience managing individual cancer patients and using HRQOL outcomes in cancer clinical trials.

Acknowledgments

This research was funded by an educational grant from AstraZeneca. We are indebted to Liz Chinchen, the librarian at the Centre for Health Economics Research and Evaluation, University of Technology, Sydney, Australia, for developing and testing our electronic search strategy, for identifying and searching all relevant online bibliographic databases, for helping identify potential source papers, and for collecting them. We would also like to thank Julia Brown, Peter Fayers, Kim Hawkins, and Galina Velikova for helpful comments about the results.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Cella D, Bullinger M, Scott C, Barofsky I; Clinical Significance Consensus Meeting Group. Group versus individual approaches to understanding the clinical significance of difference or changes in quality of life. *Mayo Clin Proc*. 2002;77:384–392.
2. Guyatt G, Osoba D, Wu AW, Wyrwich KW, Norman GR; Clinical Significance Consensus Meeting Group. Methods to explain the clinical significance of health status measures. *Mayo Clin Proc*. 2002;77:371–383.
3. Osoba D, King MT. Interpreting quality of life in individuals and groups: meaningful differences. In: Fayers PM, Hays RD, editors. *Assessing quality of life in clinical trials: Methods and practice*. 2nd ed. Oxford, UK: Oxford University Press; 2005:243–257.
4. Aaronson NK, Ahmedzai S, Bergman B, et al. The European Organization for Research and Treatment of Cancer : a quality-of-life instrument for use in international clinical trials in oncology. *J Natl Cancer Inst*. 1993;85:365–376.
5. Cella DF, Tulsky DS, Gray G, et al. The functional assessment of cancer therapy scale: development and validation of the general measure. *J Clin Oncol*. 1993;11(3):570–579.
6. Lydick E, Epstein RS. Interpretation of quality of life changes. *Qual Life Res*. 1993;2(3):221–226.
7. King MT. The interpretation of scores from the EORTC quality of life questionnaire . *Qual Life Res*. 1996;5(6):555–567.
8. King MT, Stockler MR, Cella DF, et al. Meta-analysis provides evidence-based effect sizes for a cancer-specific quality-of-life questionnaire, the FACT-G. *J Clin Epidemiol*. 2010;63(3):270–281.
9. Cohen J. *Statistical power analysis for the behavioural sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum; 1988.
10. 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*. 2003;41:582–592.
11. Landis J, Koch G. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159–174.
12. Deeks J, Higgins J, Altman D. Analysing and presenting results. In: Higgins J, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions 4.2.6. [Updated Sep 2006]. Chichester, UK: John Wiley & Sons; 2006.
13. Follman D, Elliott P, Suh II, Cutler J. Variance imputation for overviews of clinical trials with continuous response. *J Clin Epidemiol*. 1992;45(7):769–773.
14. Hedges LV, Olkin I. *Statistical Methods for Meta-analysis*. Orlando: Academic Press; 1985:110.
15. Yost KJ, Eton DT. Combining distribution- and anchor-based approaches to determine minimally important differences: the FACIT experience. *Eval Health Prof*. 2005;28(2):172–191.
16. 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):547–561.
17. Auchter RM, Scholtens D, Adak S, Wagner H, Cella DF, Mehta MP. Quality of life assessment in advanced non-small-cell lung cancer patients undergoing an accelerated radiotherapy regimen: report of ECOG study 4593. Eastern Cooperative Oncology Group. *Int J Radiat Oncol Biol Phys*. 2001;50(5):1199–1206.
18. Esper P, Mo F, Chodak G, Sinner M, Cella D, Pienta KJ. Measuring quality of life in men with prostate cancer using the functional assessment of cancer therapy-prostate instrument. *Urology*. 1997;50(6):920–928.
19. Watkins-Bruner D, Scott C, Lawton C, et al. RTOG's first quality of life study- RTOG 90-20: a phase II trial of external beam radiation with etanidazole for locally advance prostate cancer. *Int J Radiat Oncol Biol Phys*. 1995;33(4):901–906.
20. Schink JC, Weller E, Harris LS, et al. Outpatient taxol and carboplatin chemotherapy for suboptimally debulked epithelial carcinoma of the ovary results in improved quality of life: an Eastern Cooperative Oncology Group Phase II study (E2E93). *J Cancer*. 2001;7(2):155–164.
21. Demetri GD, Kris M, Wade J, Degos L, Cella D. Quality of life benefit in chemotherapy patients treated with epoetin alfa is independent of disease response or tumor type: results from a prospective community oncology study – Procrit Study group. *J Clin Oncol*. 1998;16(10):3412–3425.
22. Langer CJ, Manola J, Bernado P, et al. Cisplatin-based therapy for elderly patients with advanced non-small-cell lung cancer: implications for Eastern Cooperative Oncology Group 5592, a randomized trial. *J Natl Cancer Inst*. 2002;94(3):173–181.

23. Fallowfield L, Gagnon D, Zagari M, et al. Multivariate regression analyses of data from a randomised, double-blind, placebo-controlled study confirm quality of life benefit of epoetin alfa in patients receiving non-platinum chemotherapy. *Br J Cancer*. 2002;87(12):1341–1353.
24. Hahn EA, Glendenning GA, Sorensen MV, et al. Quality of life in patients with newly diagnosed chronic phase chronic myeloid leukemia on imatinib versus interferon alfa plus low dose cytarabine: results from the IRIS study. *J Clin Oncol*. 2003;21(11):2138–2146.
25. Eton DT, Cella D, Yost KJ, et al. A combination of distribution – and anchor-based approaches determined minimally important differences (MIDs) for four endpoints in a breast cancer scale. *J Clin Epidemiol*. 2004;57(9):898–910.
26. McQuellon RP, Thaler HT, Cella D, Moore DH. Quality of life (QOL) outcomes from a randomized trial of cisplatin versus cisplatin plus paclitaxel in advanced cervical cancer: a Gynecologic Oncology Group study. *Gynecol Oncol*. 2006;101(2):296–304.
27. Schwartz CE, Bode R, Repucci N, Becker J, Sprangers MAG, Fayers PM. The clinical significance of adaptation to changing health: a meta-analysis of response shift. *Qual Life Res*. 2006;15(9):1533–1550.
28. Klee M, King M, Machin D, Hansen H. A clinical model for quality of life assessment in cancer patients receiving chemotherapy. *Ann Oncol Adv Access*. 2000;11(1):23–30.
29. McCain NL, Zeller JM, Cella DF, Urbasski PA, Novak RM. The influence of stress management training in HIV disease. *Nurs Res*. 1996;45(4):246–253.
30. Arora NK, Gustafson DH, Hawkins RP. Impact of surgery and chemotherapy on the quality of life of younger women with breast carcinoma: a prospective study. *Cancer*. 2001;92(5):1288–1298.
31. Gustafson DH, Hawkins R, Pingree S, et al. Effect of computer support on younger women with breast cancer. *J Gen Intern Med*. 2001;16(7):435–445.
32. Velikova G, Booth L, Smith AB, et al. Measuring quality of life in routine oncology practice improves communication and patient well-being: a randomized controlled trial. *J Clin Oncol*. 2004;22(4):714–724.
33. Jaeschke RJ, Singer J, Guyatt G. Measurement of health status: ascertaining the minimally clinically important difference. *Control Clin Trials*. 1989;10:407–415.
34. Schünemann HJ, Puhan M, Goldstein R, Jaeschke R, Guyatt GH. Measurement properties and interpretability of the chronic respiratory disease questionnaire (CRQ). *J Chron Obstruct Pulmon Dis*. 2004;2(1):81–89.