The Physical Function and Social-Emotional Function Subscales of the University of Washington Quality of Life Questionnaire

Simon N. Rogers, FDS, RCS, FRCS, MD; Derek Lowe, MSc, CStat; Bevan Yueh, MD, MPH; Ernest A. Weymuller Jr, MD

Objectives: To perform a factor analysis using the University of Washington Quality of Life Questionnaire version 4 (UW-QOLv4) to establish subscales; to report their normative values and variations for patients by age, sex, extent of disease, and time from treatment; and to estimate clinical effect sizes and potential for use in comparative treatment studies.

Design: Retrospective cohort.

Setting: Regional Maxillofacial Unit, University Hospital Aintree, Liverpool, England.

Patients: Patients with primary oral and oropharyngeal cancer treated by surgery with or without adjuvant radiotherapy since 1992. A database accumulating since 1995 contains more than 2600 UW-QOLs completed by these patients. A data set of 372 patients without cancer attending 10 general dental practices provided normative data.

Main Outcome Measures: UW-QOLv4.

Results: Factor analysis indicated a 2-factor solution: (1) physical function, involving chewing, swallowing, speech, taste, saliva, and appearance, and (2) social-emotional function, involving anxiety, mood, pain, activity, recreation, and shoulder function. The best scores were for those with less advanced oral cancer tumors not requiring free-flap surgery or adjuvant radiotherapy. Older patients reported better scores, but associations were weak, and no sex differences were found. Significant differences were seen for T category, site, free-flap surgery, and adjuvant radiotherapy ($P < .001$). Preoperative scores were close to normative values. Patients regain social-emotional deficits by 1 year after surgery but continue with significant deficits in physical function. Comparative studies using these UW-QOL subscales as outcome measures should recruit at least 80 patients per treatment arm to detect moderately sized treatment effects.

Conclusion: With the UW-QOLv4, it is appropriate to analyze and report outcomes using the 2 subscales of physical and social-emotional function.

correlation coefficient ($r$). Kappa coefficient of agreement statistics ($\kappa$) measured within-patient agreement of UW-QOL subscale score quartiles in responses 1 year apart. Kappa values greater than 0.60 represent “good” agreement, and those above 0.80 “very good” agreement. An extension of this approach called weighted $\kappa$ was used to give credit for partial agreement by assigning weights to off-diagonal cells in the $4 \times 4$ agreement table. Quadratic weights were used, which base disagreement weights on the square of the amount of discrepancy, a weighting method mathematically equivalent to the intraclass correlation coefficient. The McNemar-Bowker test statistic was used to indicate systematic bias.

RESULTS

The 1992 to 2006 cohort comprised 838 patients having primary surgery for oral and oropharyngeal squamous cell carcinoma. Data were factor analyzed for 517 patients with UW-QOL data at least 9.5 months after primary surgery and closest available to 12 months (median, 15 months; interquartile range [IQR], 12-25 months). From the total cohort of 838 patients, 734 were alive at 9.5 months; thus, the 517 patients with data beyond 9.5 months represent 70.4% (517 of 734) of those alive at 9.5 months. Mean (SD) age at operation of the 517 patients was 61 (12) years, with men composing 62.7% of the sample ($n=324$). Most patients (86.1%, $n=445$) had oral cavity tumors; tumors were also located in the oropharynx (13.2%, $n=68$) and maxilla (0.8%, $n=4$). One-third (145 of 445) of the oral tumors were clinical category T3 and T4, 63.4% ($n=291$) underwent free-flap surgery, and 32.4% ($n=144$) received adjuvant radiotherapy.

FACTOR ANALYSES

Inspection of scree slopes indicated a 2-factor solution with eigenvalues above 1.0 with a very clear elbow in the plot, with the first 2 factors lying above a straight line joining factors 3 through 12. Each factor had 6 domains, and correlations among all domain scores were positive. We labeled these factors “physical” function and “social-emotional” function. Physical function involves the chewing, swallowing, speech, taste, saliva, and appearance domains; social-emotional function involves the anxiety, mood, pain, activity, recreation, and shoulder function domains. Factor 1 and 2 loadings for appearance were similar but were higher for factor 1, indicating that appearance straddles between physical and social function, with slightly more physical than social function.

Analysis of all other QOL data after 9.5 months from surgery (median, 52 months; IQR, 33-81 months) and representing a mixed bag of data from 1361 questionnaires with multiple responses from patients and data ranging up to 15 years after surgery but excluding the 517 patients first analyzed gave a similar factor structure. Analysis of QOL data for 346 patients after surgery but before 9.5 months taking the closest available data after 3 months (median, 5 months; IQR, 3-6 months) again suggested 2 factors, with the first factor composed of swallowing, chewing, speech, taste, and saliva and the second factor composed of mood, anxiety, pain, and recre-
Age and sex reference data for the UW-QOLv4 were given in showing differences between patient subgroups, as presentation, skewed slightly toward lower (worse) scores can be regarded as numerical for the purpose of least 4 domain scores are available. The 2 subscale component domain scores, with the requirement that at least one category level is not straddling between the 2.

PRESENTATION AND SCORING

Physical function and social-emotional function scores are computed as a simple average of their respective component domain scores, with the requirement that at least 4 domain scores are available. The 2 subscale scores can be regarded as numerical for the purpose of presentation, skewed slightly toward lower (worse) scores. A boxplot is an appropriate graphic to use for showing differences between patient subgroups, as given in Table 1. Results may be summarized as median (IQR) or as mean (SD).

NORMATIVE DATA

Age and sex reference data for the UW-QOLv4 were collected from 372 patients attending 10 general dental practices. No obvious differences were noted in physical function and social-emotional function scores by age and sex. Overall median (IQR) normative scores were 100 (95-100) for physical function and 90 (74-100) for social-emotional function. Mean (SD) scores were 95 (10) for physical function and 83 (19) for social-emotional function.

PATIENT CHARACTERISTICS AND VARIATIONS IN SUBSCALE SCORES

Overall median (IQR) patient scores at approximately 1 to 2 years were 73 (59-90) for physical function and 80 (61-92) for social-emotional function. Mean (SD) scores were 71 (21) for physical function and 74 (20) for social-emotional function. In relation to normative data, the physical deficit was much more pronounced 1 to 2 years after surgery than was the social-emotional deficit.

The best set of physical and social-emotional scores was seen for patients with less advanced oral cancer tumors not requiring free-flap surgery or adjuvant radiotherapy (Table 1). This group had mean scores of 87 for physical function and 84 for social-emotional function, the latter being similar to normative levels. Physical function scores were lower for patients undergoing free-flap surgery with adjuvant radiotherapy irrespective of tumor category (P < .001 for T1-T2 and for T3-T4), but social-emotional function scores were more similarly distributed (P = .58 for T1 and T2 and P = .06 for T3 and T4).

Older patients reported better physical function and social-emotional function scores, but the associations were weak (Spearman r = 0.09 and r = 0.11 respectively; P = .03 for both). No significant differences were noted for sex. Significant differences were seen in both subscale scores for T category (T1 and T2 vs T3 and T4, P < .001), site (oral vs oropharyngeal, P < .01), free-flap surgery (yes vs no, P < .001), and adjuvant radiotherapy (yes vs no, P < .001). Reference data for patients 1 to 2 years after primary surgery are given in Table 1 according to different patient characteristics.

TEMPORAL VARIATION

We previously reported deficits in UW-QOLv4 domain scores after treatment, with some degree of recovery seen to approximately 12 months but little change in mean scores thereafter. Similar trends can be seen for the physical and social-emotional function scores, more notably the deficits in physical functioning (Table 2). Although the preoperative data are limited, the subscale mean scores are close to the normative values. Table 2 provides trends across time and suggests that patients just about regain any social-emotional deficits by 1 year after surgery to levels not that dissimilar to the normative levels. In contrast, patients continue to have significant residual deficits in physical function into the long term. These observations, however, could mask the more subtle changes that can be observed from the full profile of results (Figure). For physical function, the direction of change from the time of surgery is the same for each component domain, and the subscale score reflects this (Figure, A). For social-emotional function, the immediate deficits in shoulder function, activity, and recreation are counterbalanced by improvements in pain and

<table>
<thead>
<tr>
<th>Subscale</th>
<th>Total (N=517)</th>
<th>T1-T2, No Flap, No RT (n=135)</th>
<th>T1-T2, Flap, No RT (n=96)</th>
<th>T1-T2, Flap, RT (n=51)</th>
<th>T3-T4, Flap, No RT (n=58)</th>
<th>T3-T4, Flap, RT (n=78)</th>
<th>Oral Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical function score</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>74 (20)</td>
<td>87 (15)</td>
<td>74 (20)</td>
<td>72 (22)</td>
<td>73 (18)</td>
<td>65 (23)</td>
<td>70 (22)</td>
</tr>
<tr>
<td>Median (IQR)</td>
<td>80 (61-92)</td>
<td>87 (75-95)</td>
<td>76 (61-92)</td>
<td>79 (55-91)</td>
<td>77 (61-88)</td>
<td>68 (49-83)</td>
<td>74 (51-88)</td>
</tr>
<tr>
<td>Social-emotional function score</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>74 (20)</td>
<td>87 (15)</td>
<td>74 (20)</td>
<td>72 (22)</td>
<td>73 (18)</td>
<td>65 (23)</td>
<td>70 (22)</td>
</tr>
<tr>
<td>Median (IQR)</td>
<td>80 (61-92)</td>
<td>87 (75-95)</td>
<td>76 (61-92)</td>
<td>79 (55-91)</td>
<td>77 (61-88)</td>
<td>68 (49-83)</td>
<td>74 (51-88)</td>
</tr>
</tbody>
</table>

Abbreviations: IQR, interquartile range; RT, radiotherapy; UW-QOL, University of Washington Quality of Life Questionnaire.

a Median of 15 months; IQR, 12-25 months.

b A small residual number of 32 (517−485) patients with oral cancer who had other combinations of T category/flap status/radiotherapy status are not included in the table.

Table 1. Reference Data for the UW-QOL Physical Function and Social-Emotional Function Scores 1 to 2 Years After Surgery for Key Patient Groups

Figure

©2010 American Medical Association. All rights reserved.
Table 2. Physical Function and Social-Emotional Function Scores* by Time From Surgery for Patients With Oral and Oropharyngeal Cancer

<table>
<thead>
<tr>
<th>Patient Groupb</th>
<th>Patients, No.</th>
<th>Physical Function</th>
<th>Social-Emotional Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preoperative UW-QOL</td>
<td>267</td>
<td>88 (80-100)</td>
<td>81 (69-89)</td>
</tr>
<tr>
<td>UW-QOL closest to 3 mo after operation but before 9.5 mo</td>
<td>343</td>
<td>66 (54-80)</td>
<td>73 (55-87)</td>
</tr>
<tr>
<td>UW-QOL closest to 12 mo but after 9.5 mo</td>
<td>505</td>
<td>73 (59-90)</td>
<td>80 (61-92)</td>
</tr>
<tr>
<td>All other UW-QOL data after 9.5 mo</td>
<td>1346</td>
<td>73 (60-91)</td>
<td>83 (66-92)</td>
</tr>
</tbody>
</table>

Abbreviations: IQR, interquartile range; UW-QOL, University of Washington Quality of Life Questionnaire.

*Range, 0-100.

The first 3 groups comprise different sets of patients; but some patients are in more than 1 group, although patients are represented only once in each group. The last group comprises a mixed bag of questionnaire data with multiple responses from patients ranging from 9.5 months to 15 years after surgery but excluding the patients in the third time group (ie, closest to 12 months).

ASSOCIATION WITH OTHER UW-QOL GLOBAL MEASURES

In the 517 patients with UW-QOL data at least 9.5 months after primary surgery and closest available to 12 months, both global QOL measures (health-related and overall QOL) were more strongly correlated with social-emotional function (Spearman r=0.70 and r=0.70, respectively) than with physical function (r=0.47 and r=0.42, respectively). This stronger association with social-emotional scores was also seen in analyses of all other QOL data late after surgery (median, 52 months) and early after surgery (median, 5 months). For the 517 patients, the Spearman correlation between the 2 global measures of QOL was r=0.82, whereas the correlation between the 2 subscales was r=0.63.

INTERPRETABILITY, CLINICAL EFFECT, AND SAMPLE SIZE

To use a subscale score as an outcome measure requires a definition of what is meant by a meaningful clinical change across time. One approach is effect size obtained by dividing the mean change in a group of patients by the standard deviation in the prechange data. A small effect size represents approximately 0.20 of a standard deviation, a moderate effect size approximately 0.50, and a large effect size approximately 0.80. Results at 1 to 2 years gave subscale SDs of 20 (Table 2). Thus, for patients recruited at this time, a mean change across time
of 4 U in the subscale score would be a small change, 10 U a moderate change, and 16 U a large change. Results before treatment gave subscale SDs of 15. Therefore, for patients recruited before treatment, a mean change across time of 3 U would be a small change, 7.5 U a moderate change, and 12 U a large change.

Standard statistical methods for computing sample size for a comparative study (with an α of 0.05 and study power of 80%) indicate that approximately 400 subjects per group are needed in analysis to detect small treatment effects, 64 per group to detect moderate effects, and 26 per group to detect large effects. Adjustments in numbers recruited are needed to allow for loss to follow-up, including death, so as to attain these numbers for analysis.

COMPOSITE SCORE

These data suggest that 2 subscales are more appropriate than a single composite 12 score. Some researchers are accustomed to using the composite score, so some summary statistics are given herein. Because the 2 subscale scores are each an average of 6 domain scores, the single composite 12 score is, in effect, the mean of the 2 subscale scores. We commented on the trend for patients to have a more significant physical deficit after treatment and persisting into the longer-term, whereas the social-emotional scores recover to just about baseline values. The composite 12 score would fall between these 2 trends, showing a deficit and some recovery but not to baseline values. Median (IQR) normative scores were 94 (85-98), with mean (SD) of 89 (13). In the 517 patients with UW-QOL data at least 9.5 months after primary surgery and closest available to 12 months, Spearman correlation with the physical function subscale was r = 0.90, with the social-emotional function subscale was r = 0.89, with global health-related QOL was r = 0.62, and with global overall QOL was r = 0.59. Median (IQR) composite score was 76 (60-88), and mean (SD) value was 73 (19).

COMMENT

This article presented data on the UW-QOLv4 for factor analyses, presentation and scoring of the resulting subscales, their comparison with normative data, patient characteristics and variations in subscale scores, temporal variation, association with other UW-QOL global measures, interpretability, clinical effect, and sample size. This article introduced 2 new UW-QOL subscales, which we called physical function and social-emotional function to reflect their component items. Each subscale score is computed as the simple average of its 6 components and is scaled from 0 to 100. The ultimate test for these scales is whether they work well in practice to discriminate between groups and to respond to changes in clinical condition across time. The results so far suggest that they do.

The data set was of sufficient size to perform both factor analyses and to make comparisons by patient characteristics, across time from treatment, and for normative reference data. It is recognized that there are limited pretreatment data (one-third of patients only), and the analyses was limited to primary surgery cases in one particular group of patients with head and neck cancer. However, in exploring the psychometric properties of the UW-QOLv4, the findings from this cohort should be applicable to other patient groups and does not reduce the bearing of the findings.

Disease-specific instruments have several advantages. They reduce patient questionnaire burden and increase acceptability by including only relevant dimensions. This may increase responsiveness. Whether dimensions should be combined in a summary score depends on whether such a score is useful and interpretable and whether there are any interactions between treatment and dimensions (ie, cancelling out of effects). Because the UW-QOLv4 domains and global scales are categorical, with only a few discrete scoring options and a skewed response, the new physical function and social-emotional function scales, with their many possible scores, should increase responsiveness and precision. The 2 subscale scores have face and content validity. It was remarkable that the shoulder domain hardly loaded to the physical function factor. The fact that the shoulder domain is in the social-emotional function subscale likely reflects the wording of the domain for work and hobbies. Rather than physical shoulder function, the question seems to pertain more to the social consequences of shoulder problems. The 2 subscales of the UW-QOLv4 have a similar construct as other questionnaires, such as the Medical Outcomes Survey 36-Item Short Form Health Survey, which has physical and mental component scores. The 2 constructs are different, and it is recognized that patients can go back to baseline emotional functioning despite physical limitations.

The 2 subscales each have 6 domains, and this gives better discrimination than any single domain. It is accepted that we can never be sure that using equal weights for scaling the subscale scores (ie, the simple average) gives the best estimate of physical function or social-emotional function. However, there will be further validation as colleagues report outcomes using the UW-QOLv4 as to how well equal-weight scaling works in practice.

The present data support the stability of the subscale scores from approximately 1 year after treatment and onward, their validity, and their responsiveness (sensitivity to clinically significant changes across time). The high κ and intraclass correlation values between repeated applications 12 months apart in the absence of clinical treatment or changes in disease status are remarkably high, suggesting that there is high test-retest reliability when conditions are stable beyond 1 year. The median difference in score of 0, the tight IQR of within ±5 U for differences, and the lack of systematic error in repeated applications also enhance the measure. Aspects of validity are explored by contrasting normative with patient data, within patients across time and between subgroups of patients with differing clinical characteristics.

One of the most important areas for further development is in making quantitative change scores for QOL more clinically meaningful. Ringash et al defined minimal important difference (MID) as the smallest difference that reflects a clinically important change in
score and noted that having some knowledge of MID is important for determining sample size when planning trials with QOL end points. Most published MID estimates fall in the range 5% to 10% of the instrument range. The present results are consistent with the broad conclusion of Ringash et al\(^2\) regarding MID as 5% to 10% of the instrument range. In the present study, for patients recruited before treatment, a mean change of 7.5 U constituted a moderate effect (50% of SD), whereas for patients recruited approximately 1 year after treatment, a mean change of 10 U constituted a moderate effect.

However imprecise these percentages might seem, sample size calculations are important because they indicate the right order of magnitude of size of the study to be recruited. In clinical research, most advances are in small steps, and many effects, if real, are likely to be small to medium and are unlikely to be large. This indicates that 160 (80 per group) should be regarded as the minimum requirement for recruitment to a 2-armed randomized controlled trial to be able to detect moderately sized differences and allowing for 20% patient attrition.

In conclusion, the UW-QOLv4 is brief and simple to complete. It had minimum patient burden. Despite its brevity, the questionnaire does have psychometric validity. The identification of 2 subscales, physical function and social-emotional function, potentially increases its responsiveness and precision. They also allow a realistic estimate of sample size for clinical trials. They are to be preferred to the single aggregate composite 12 score. Questionnaire analyses and reporting should include both the physical function and social-emotional function subscales.

Submitted for Publication: January 11, 2009; final revision received May 4, 2009; accepted May 14, 2009. Correspondence: Simon N. Rogers, FDS, RCS, FRCS, MD, Regional Maxillofacial Unit, Aintree University Hospital’s National Health Service Foundation Trust, Aintree, Liverpool L9 7AL, England (snrogers@doctors.org.uk). Author Contributions: All authors had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Study concept and design: Rogers, Lowe, and Yueh. Acquisition of data: Rogers. Analysis and interpretation of data: Rogers, Lowe, Yueh, and Weymuller. Drafting of the manuscript: Rogers, Lowe, and Yueh. Critical revision of the manuscript for important intellectual content: Rogers, Yueh, and Weymuller. Statistical analysis: Lowe and Yueh. Obtained funding: Rogers. Study supervision: Rogers. Financial Disclosure: None reported.

REFERENCES