Quality-of-Life Outcomes in the Evaluation of Head and Neck Cancer Treatments

Seth Schwartz, MD; Donald L. Patrick, PhD, MSPH; Bevan Yueh, MD

Objectives: To review the published literature to evaluate the design, use of terminology, and interpretation of results in studies using quality-of-life (QOL) instruments to measure differences between head and neck cancer treatments at a point in time or to report changes over time in one or more treatment groups.

Data Source: MEDLINE search for subject headings “head and neck neoplasms” (as a main topic) and “quality of life” or “health status” restricted to English-language sources and a 10-year period from 1989 to 1999.

Study Selection: Four hundred forty-five abstracts were reviewed to find articles using an instrument to compare head and neck cancer therapy groups with a QOL outcome (13.7% included).

Data Extraction: Two readers reviewed each article to determine how terminology was used, if a scientific study design was used, and if differences or changes in scores were clinically interpreted.

Results: Sixty-one articles were reviewed. Forty different instruments were used. Terminology was used inconsistently in 21 (34.4%) of the 61 articles. A scientific study design was used in only 11 (18.0%) of the 61 articles (P<.001). A clinical interpretation of results was given in 16 (26.2%) of the 61 articles (P<.001).

Conclusions: While QOL outcomes show promise for assisting with treatment decisions in head and neck cancer therapy, few studies using instruments to measure QOL outcomes are hypothesis driven and clinical interpretations of results are not commonly provided. We recommend that future studies identify the construct to be measured, specify comparator groups and hypotheses a priori, and provide clinical interpretations of results.


HEAD AND NECK cancer represents about 5% of cancers diagnosed annually. Head and neck cancer generally refers to squamous cell carcinoma of the nasopharynx, oral cavity, oropharynx, hypopharynx, or larynx. Surgery and radiotherapy are the 2 primary modes of therapy for head and neck cancer, although chemotherapy is increasingly used for advanced disease. Despite intensive work in developing new treatment alternatives, these treatments have, thus far, had little influence on survival in most subsites of the head and neck. Since 1989, there have been more than 300 articles in the head and neck literature that refer specifically to QOL. However, there is wide variation in what is meant by “quality of life.” Authors have used QOL to refer to a variety of distinct concepts, such as symptoms alone, functional status, health status, health-related quality of life (HRQOL), and overall (global) QOL. Global QOL encompasses a patient’s social; emotional; psychological; and physical functional status, symptoms such as pain; and environmental factors such as income, opportunities, family support, and lifestyle choices. These individual components are referred to as “domains.” Because factors such as income and family support are often unaffected by treatment, clinicians generally focus on HRQOL, which refers to the aspects of QOL pertaining to a health or medical concern.
MATERIALS AND METHODS
SEARCH STRATEGY

A MEDLINE search was performed through the National Library of Medicine's Internet site “Internet Grateful Med.” Search terms were “head and neck neoplasm” as a main topic and “quality of life” or “health status.” The search was then limited to the English-language sources and to a 10-year period from 1989 to 1999. This search strategy resulted in 445 articles.

Part 1: Exclusion and Classification of Articles

The 445 abstracts were then reviewed and classified by content and study design to identify those addressing QOL and antineoplastic treatments (including reconstruction) for squamous cell carcinoma of the nasopharynx, oral cavity, oropharynx, hypopharynx, or larynx. One of us (S.S.) reviewed all abstracts to identify articles that met inclusion criteria, and the 2 additional investigators (D.L.P. and B.Y.) each reviewed the same 10% of the articles chosen at random to assess correlation (k=0.89, P<.001). Articles about the esophagus (n=113), the thyroid gland (n=17), or other topics not specifically head and neck cancer (n=19, Table 1) were excluded. Twelve articles were excluded because they reported on interventions that were not cancer treatments (ie, nutritional interventions, psychosocial interventions, screening, or smoking cessation interventions). Twenty-one articles that studied only nutritional status, not QOL, were excluded. Sixty-one articles reported studies in which QOL was not assessed (an article was classified as assessing QOL if either it expressed intent to assess QOL or it reported the effect of a treatment on QOL in the abstract).

For broad discussions about QOL, the failure to draw distinctions between the various domains of QOL is usually not problematic. However, precise scientific study of these issues rely on psychometric questionnaires (instruments) that target a specific construct such as HrQOL or functional status. When authors do not specify what construct they are measuring, the meaning of the results can be ambiguous and the appropriateness of the chosen instrument can be difficult to determine. In this article, we will use the term QOL to refer to all QOL–related constructs that were obtained through the use of questionnaires, and use the term HrQOL to distinguish it from other domains such as functional status.

Most instruments report outcomes as numerical scores. Studies use these scores to compare 2 or more treatment groups at a point in time (discriminative) or 1 or more groups receiving the same treatment over time (evaluative). Instruments may prove sensitive to differences in QOL scores between groups or changes over time, but interpreting what a given change score or difference in scores means is still a difficult task. In other words, finding a statistically significant difference between treatment groups does not guarantee that the difference is clinically relevant. Specifically, reported results may not be meaningful to clinicians or patients hoping to gain guidance from published studies. Standards for reporting results in an interpretable manner are evolving rapidly.

Several studies exist that evaluate the validity and method of QOL measurement in the general medical literature, but, to our knowledge, no studies have investigated how well instruments are being used to measure QOL outcomes in head and neck cancer. This study evaluated how discriminative and evaluative differences are studied and reported in the head and neck literature and evaluated whether these differences have been clinically interpreted. To do this we reviewed all discriminative and evaluative studies on QOL in head and neck cancer identified through MEDLINE. We hypothesized that there would be many studies that were inaccurate in their use of terminology describing QOL outcomes, failed to use hypothesis-driven study designs, and failed to interpret the clinical significance of reported differences.

RESULTS

A total of 98 articles were grouped and classified. Thirty-three articles (33.7%) were descriptive studies in which a case or case series was presented or a description of a single group of patients was presented leaving 65 studies (66.3%) that made comparisons. Sixty-one (93.8%) articles reported the effect of a treatment on QOL in the abstract (REPRINTED) ARCH OTOLARYNGOL HEAD NECK SURG/VOL 127, JUNE 2001 WWW.ARCHOTO.COM

©2001 American Medical Association. All rights reserved.
at least referenced the study in which the psychometric properties of the chosen instrument were described.

Was a Scientific Study Design Used?

We asked if comparator groups were specified a priori and a formal hypothesis was tested. All clinical studies should have an explicit hypothesis to test because this increases the clinical plausibility and decreases the possibility that observed differences resulted by chance.68 For this review, the investigators had to state explicitly which group they expected to have a better outcome and present this hypothesis in the “Introduction” or “Methods” section of the article to be credited. To be considered a priori, the comparator groups had to be explicitly defined and the comparison between them considered a goal of the study. Post hoc comparisons or partitioning of the data were not credited.

Was an Effort Made to Interpret Differences in QOL Scores?

This consideration primarily applied to articles that used instruments with composite scores or composite domain scores. Composite scores are obtained by combining the results of multiple questions into a single score. Composite scores can be difficult to interpret because they do not reflect an answer to any one question. To determine if investigators interpreted score differences or changes in scores over time, we used criteria established by the Scientific Advisory Committee (SAC) of the Medical Outcomes Trust.70 Scientific Advisory Committee of the Medical Outcomes Trust recently compiled recommendations for the clinical interpretation of QOL data.71 They advocated 2 principal methods. The first is through use of an explicitly stated minimally important difference. This means that the investigation must state a priori how big a difference there must be between groups to be clinically meaningful. The second is by using familiar clinical anchors. This means that QOL scores are related to clinically familiar concepts like community norms, recognized clinical conditions, recognized life events, or meaningful adjectives like “better” and “worse.” We used these guidelines to perform this review.34

While a discussion of the importance of interpretability has been present in the health outcomes literature for years, these guidelines were not published at the time several of the earlier articles were published. Therefore, we were generous in giving credit for fulfilling these guidelines. Any attempt to use these criteria was credited as achieving interpretability. In addition, credit was given if 2 readers (D.L.P. and B.Y.) concurred that the results were clinically interpretable. For example, a single-item instrument with adjectives like “worse,” “same,” and “better” as potential answers was usually interpretable.

DATA COLLECTION AND ANALYSIS

A form was developed to standardize data extraction from the articles. Ten percent of the 61 articles were selected at random to test the extraction questionnaire. Each of the 3 reviewers read and extracted data from the selected articles. The judgments of each reviewer were then compared and discussed. Minor corrections to the data extraction form were made to ensure clarity and accuracy of data extraction. All 61 articles were then reviewed again. One investigator read all of the articles (S.S.) while the other 2 investigators (D.L.P. and B.Y.) split the articles. Each article was, therefore, reviewed by 2 investigators. Data extraction results were then compared for each article. Disputes were resolved by consensus. Data were analyzed using a test of binomial proportions in Statistical Product and Service Solutions (SPSS Inc, Chicago, Ill), a statistical software package.

of the 65 comparative studies used an instrument. These 61 articles were the subject of part 2 of this study.

We then classified the types of comparisons made in each of the 61 articles. Several studies contained more than one form of comparison. Twenty-eight articles (45.9%) compared 2 or more groups receiving different cancer therapies. Nineteen (31.1%) of the remaining articles compared scores at 2 or more distinct points in time (ie, before and after treatment), but did not separate groups by type of cancer treatment. The remaining 14 studies (23.0%) compared scores in 2 or more groups defined by patient, demographic, or disease characteristics, but not by treatment type or over time.

Overall, 40 different instruments were used. Of these, 30 were previously validated. The EORTC QLQ-C30 (European Organization for Research and Treatment of Cancer Quality of Life Questionnaire 30) was the most commonly used instrument (n=15). The head and neck module of the EORTC was used in 13 articles. The University of Washington Quality of Life questionnaire was the next most commonly used instrument (n=9). In 10 articles (16.4%), at least one instrument was used inappropriately. In most circumstances, this means that an instrument designed to measure functional status or performance status was erroneously used to measure overall QOL.

Throughout the articles, 16 different terms were used to classify QOL outcomes. The general term “quality of life” was the most commonly used term. It was used in 38 articles (62.3%). Functional status and HrQOL were the next most common terms. They were used in 16 (26.2%) and 7 (11.5%) of the articles, respectively. Other commonly used terms included “health status,” “life satisfaction,” “well-being,” “utility,” “performance status,” and “psychological function.”

Inaccurate usage of terminology was also most common with QOL. In 17 (44.7%) of 38 articles, the usage was inaccurate. Inaccuracies usually resulted from failing to recognize distinctions between functional status and overall QOL. The terms health status and well-being were used synonymously with quality of life. The same inaccuracies occurred with these terms. The other term not used appropriately was performance status, which was also equated to QOL in one article. Overall, 21 (34.4%) of the articles used at least 1 term inaccurately.

Few articles used a proper scientific study design. Fifty-two articles (85.2%, P<.001) specified comparator groups a priori. Only 11 (18%) of the 61 articles (P=.001) stated a testable hypothesis.

Only 16 articles (26.2%) had some form of clinical interpretation of results (P<.001, Table 2). Six articles
addressed neither the statistical nor the clinical implication of their findings. Clinical significance was most commonly established using interpretive anchors such as community norms or classifying adjectives like better, worse, and same (14 of 16 articles). Three of the articles specified what a minimally important difference in scores was. For example, several of the articles using the 36-Item Short-Form Health Survey specified that a 10-point difference in score was clinically significant. There were no other notable methods used for interpreting between group mean score differences. Prior to 1993 there were no articles with an attempt at interpretation. After 1993, more articles were identified per year and more made efforts at interpretation. The proportion of articles with interpretation, however, did not increase significantly from 1994 to 1999.

The goal of QOL measurement in this field is primarily to make informed decisions between treatment alternatives. Reliably measuring QOL outcomes and reporting differences in a clinically meaningful way is critical. Studies that use a validated instrument to measure QOL scores for a treatment group, compare those results to QOL scores for another treatment group, and report the difference in a meaningful fashion have the potential to aid clinicians and patients who are trying to decide between competing treatment strategies.

It is in this context that this review of the head and neck literature was undertaken. Sixty-one (93.8%) of the articles that intended to measure discriminative or evaluative differences used an instrument. While there are no data to compare the performance of the head and neck literature with that of other medical specialties, 93.8% represents widespread recognition of the need to explicitly measure QOL.

In total, 40 different instruments were used in the 61 articles that we reviewed. Ten unvalidated instruments were used and 10 were inappropriately used. This means that these articles reported different outcomes than were actually measured or they reported outcomes that may not have been measured accurately.

Basic to the idea of interpretability is the concept of familiarity. Once clinicians become familiar with a measure through repeated applications, the measure tends to become more meaningful. A frequently used example to explain this concept is the measurement of blood pressure. High blood pressure is defined as a systolic pressure of 140 mm Hg or greater. There is nothing inherently meaningful about the number 140 mm Hg, but this cutoff point has meaning because it is so commonly used. Ideally, QOL measurement would achieve similar familiarity. With so many different instruments in use, many of which purport to measure the same things, it is nearly impossible for practitioners to become familiar with scoring norms and clinically meaningful cutoff points in all of the instruments. Narrowing the number of instruments in use would enhance the meaning of results from QOL outcome studies, but this is impracticable as no instrument has come to represent a standard criterion and different instruments measure different domains.

There is no reliable or acceptable way to determine how independent domains should be combined to provide an overall QOL score. The complexity of these relationships lends to the confusion surrounding what is meant by QOL. Therefore, it is important to use terminology accurately. We found 16 different terms used to label QOL outcomes. Quality of life was the most common. It was also the term most frequently used inappropriately or inaccurately. With one exception, the other terms used inaccurately also involved failure to distinguish specific domains from general catchall phrases.

A principal component of scientific evaluation is hypothesis testing. Only 11 studies (18.0%) stated an explicit hypothesis. These results support our own hypothesis that many QOL studies are not hypothesis driven. Some investigators expressed intent to compare groups or evaluate treatments, but they did not state which group they expected to have a superior outcome. Some even commented that their results either met with or contradicted their expectation, but they did not state these expectations in their introduction or experimental design. These articles were not credited here as having a hypothesis. Comparative studies using QOL measures as a basis for comparison are both more valid and easier to understand if an explicitly stated hypothesis is tested.

Another scientific issue was potentially biased conclusions from missing data. For example, patients with recurrent disease were excluded from evaluation in 19 (35.2%) of 54 articles to which this applies (the time frame of some of the studies was too short for disease to recur).

Table 1. Reasons for Exclusions of Articles*  

<table>
<thead>
<tr>
<th>Reasons for Exclusion</th>
<th>No. of Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>445</td>
</tr>
<tr>
<td>Esophagus</td>
<td>113</td>
</tr>
<tr>
<td>Thyroid gland</td>
<td>17</td>
</tr>
<tr>
<td>Other site</td>
<td>8</td>
</tr>
<tr>
<td>Benign disease</td>
<td>7</td>
</tr>
<tr>
<td>Other histology</td>
<td>4</td>
</tr>
<tr>
<td>Not cancer treatment</td>
<td>12</td>
</tr>
<tr>
<td>Nutrition (not QOL)</td>
<td>21</td>
</tr>
<tr>
<td>QOL not assessed</td>
<td>61</td>
</tr>
<tr>
<td>No primary data</td>
<td>82</td>
</tr>
<tr>
<td>Validation</td>
<td>14</td>
</tr>
<tr>
<td>Instrument comparison</td>
<td>2</td>
</tr>
<tr>
<td>Patient vs caregiver rating</td>
<td>5</td>
</tr>
<tr>
<td>Not human</td>
<td>1</td>
</tr>
<tr>
<td>Total remaining</td>
<td>98</td>
</tr>
</tbody>
</table>

*QOL indicates quality of life.

Table 2. Percentage of the 61 Studies Using Method for Establishing Significance

<table>
<thead>
<tr>
<th>Type of Significance</th>
<th>Percentage of Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical</td>
<td>90.1</td>
</tr>
<tr>
<td>Clinical</td>
<td>26.7</td>
</tr>
<tr>
<td>Anchors</td>
<td>21.7</td>
</tr>
<tr>
<td>Minimally important difference</td>
<td>3.3</td>
</tr>
<tr>
<td>Both</td>
<td>1.7</td>
</tr>
</tbody>
</table>
These patients are likely doing worse than those without recurrence. Excluding them from an analysis results in artificially improved QOL scores. If disease recurred in equal percentages of each comparator group, this may not bias the comparison, but routine exclusion of patients with ongoing disease from QOL assessment misrepresents the actual burden of QOL lost or gained.

All of the studies included in this review compared the QOL scores of multiple, distinct groups or of one group over time. In each instance, mean scores were reported and the authors drew a conclusion about which group had better HrQOL (or was superior in some independent domain) based on these scores. According to the guidelines explained above, only 16 investigators (26.2%) made some effort to interpret the clinical meaning of the difference that was measured. These results support our second hypothesis that many current QOL studies do not make an attempt to interpret results. While articles stressing the importance of clinical interpretability have been published in the literature for over 10 years, it is only in the last few years that more general acceptance of these concepts became widespread. Consequently, we compared the use of interpretation by year of publication. While none of the articles published prior to 1993 interpreted results, 5 (38.5%) of 13 articles published in 1999 used some method to interpret the clinical meaning of the results. In recent years, however, the percentage of articles with interpretation has not increased markedly. In this review, our bias was toward giving credit for interpretation if any effort was made to do so. Strictly speaking, several of the articles to which we gave credit fall short of actually meeting full criteria for interpretability. Future investigators may consider correlating QOL outcomes to currently meaningful anchors like performance status indices, time to recurrence, or weight loss.

There are major obstacles to performing research on QOL outcomes in general. These can be divided into 2 major categories: practical and methodological. Practical considerations include the amount of time and resources required to collect and analyze large amounts of questionnaire data, the difficulty in selecting an appropriate instrument, and, as mentioned above, the bias created by missing data on those most severely affected by their disease. Methodological considerations include the inability of instruments to capture patient’s adaptation to the conditions of their disease state, the difficulties of comparing outcomes across populations, and, as we emphasize in this article, the lack of interpretation of what QOL score differences mean in clinical practice.

Despite these barriers, the importance of using QOL outcomes in evaluating head and neck cancer treatment has grown because it shows promise as a means of deciding between treatments when no survival advantage is afforded by one modality over another. However, if the results of QOL studies are not made meaningful to clinicians, the potential of these measures will not be realized. Therefore, we make the following recommendations. Studies using instruments to measure QOL outcomes and to compare groups should be precise in their identification of what construct is to be measured, for example, HrQOL, functional status, or depression. Comparator groups should be identified a priori and a testable hypothesis should be presented. Finally, the clinical importance of reported differences must be interpreted using familiar anchors or minimally important differences until the instruments are more widely used and their scores understood.

Accepted for publication January 18, 2001.

This investigation was supported in part by Basic Sciences Training in Otolaryngology grant DC00018 from the National Institutes of Health, Bethesda, Md (Dr Schwartz), and by career development award CD-98318 from the Department of Veterans Affairs, Veterans Health Administration, Health Services Research and Development Service, Washington, DC (Dr Yuch). The views expressed in this article are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs.

Corresponding author and reprint: Seth Schwartz, MD, Otolaryngology—Head and Neck Surgery, University of Washington Medical Center, 1959 NE Pacific St, Box 356515, Seattle, WA 98195-6515 (e-mail: schwarsr@u.washington.edu).

REFERENCES