NUMERACY AND THE SHORTCOMINGS OF UTILITY ASSESSMENT IN HEAD AND NECK CANCER PATIENTS
Seth R. Schwartz, MD, MPH,1 Jennifer McDowell, MS,2 Bevan Yueh, MD, MPH1,2,3

1 Department of Otolaryngology/Head and Neck Surgery, University of Washington Medical Center, 1959 N E Pacific St., Box 356515, Seattle, Washington 98195-6515. E-mail: schwarsr@u.washington.edu
2 Health Service Research & Development Service, VA Puget Sound Health Care System, Seattle, Washington
3 Department of Health Services, University of Washington School of Public Health and Community Medicine, Seattle, Washington

Abstract: Background. Because survival differences between surgical and nonsurgical treatment for head and neck cancer (HNC) are hard to detect, increasing focus has been placed on quality of life (QOL) differences after treatment. Utility assessment provides insight into QOL. Evidence suggests that a patient’s comfort with numerical concepts (“numeracy”) may influence utility measures. We hypothesize that patients who are nonnumerate provide inconsistent utility data in QOL studies.

Methods. New HNC (n = 18) patients were recruited to participate. Patients completed a numeracy questionnaire, a utility assessment, and a global QOL questionnaire. Higher scores reflect better function. Interviewers rated the functional level of each patient. For both numerate and nonnumerate patients, utility scores were compared with global QOL (good vs poor) and observer-rated function.

Results. Half of the patients were numerate. Numerate patients who rated their QOL as good had significantly higher utility scores than did patients with poor global QOL (0.95 vs 0.43, p = .03). In contrast, nonnumerate patients with good QOL had lower utility scores than did patients with poor QOL (0.45 vs 0.77, NS). Utility scores for numerate patients correlated well with observer-rated function (r = 0.41 to r = 0.57), whereas those of nonnumerate patients did not (r = −0.16 to r = 0.06).

Conclusions. QOL evaluation through utility assessment may provide inaccurate and contradictory data about patient functioning for nonnumerate patients. This may confound QOL assessment when interpreting utility data.

Keywords: utility; numeracy; head and neck cancer; quality of life; function

Because surgical and nonsurgical treatments for head and neck cancer often lead to equivalent survival rates, assessment of alternative outcomes such as quality of life (QOL) is of growing importance. Prior studies have provided useful insight into the impact of cancer treatment on psychological, physical, emotional, and social domains. They have not, however, been able to determine which treatment leads to better overall, global QOL.

As
a result, physicians and patients have difficulty selecting a preferred treatment.

Recent studies have shown that patients’ quantitative skills and comfort with numerical expressions (numeracy) may impair their ability to provide meaningful responses to QOL assessment measured with utility instruments. A utility score is a quantified expression of preference for a specific health state and can be thought of as the value patients assign to their state of health or improvement in health status. Utility scores represent a global measure of QOL and help to ascertain patient preferences for various health states. Utility scores range from 0 to 1, with 0 typically representing the worst possible health state (for example, death), and 1 the best possible outcome (for example, perfect health). In general, subjects less bothered by their symptoms have higher utility scores than do subjects who are more troubled by their symptoms. Therefore, subjects with similar symptoms may have different utility scores because they have differing levels of bother.

Existing studies have shown that, even in educated samples, less than one third of patients demonstrated comfort with numerical expressions. If two thirds of patients in a given study provide invalid responses to the measurement tool, any real differences between treatment groups might be obscured.

We hypothesized that only a fraction of the patients with head and neck cancer are comfortable with numerical expressions and that this numerate group would provide meaningful responses to utility assessments, whereas nonnumerate patients would not. Furthermore, we hypothesized that pooling the data from numerate and nonnumerate patients would obscure differences in utility between groups with varying QOL.

The purpose of this study was, therefore, to determine whether lack of numeracy influenced patients’ ability to provide meaningful QOL data measured as utility in a cohort of head and neck cancer patients.

METHODS

Patient Population. We enrolled new patients with squamous cell carcinoma of the head and neck who were seen at the Otolaryngology-Head and Neck Surgery clinics at the University of Washington and the Seattle VA Puget Sound Medical Center Otolaryngology and Radiation Oncology clinics from July though November 2001. We excluded patients with prior treatment for head and neck cancer. Patients who did not speak English or were otherwise unable to complete the necessary questionnaires were also excluded. Finally, patients who were expected to seek treatment and follow-up in the community were excluded to maximize follow-up.

Patients were recruited before initiation of treatment. Patients agreeing to participate underwent a semistructured interview. This interview consisted of an objective assessment of patient function by use of the Performance Status Scale (PSS). Patients then completed the global QOL question from the University of Washington Quality of Life questionnaire (UWQOL), a utility exercise and a test of numeracy. These are described below.

Global Quality of Life Question. This was a single-item question with six possible responses that stated:

Overall quality of life includes not only physical and mental health, but also many other factors, such as family, friends, spirituality, or personal leisure activities that are important to your enjoyment of life. Considering everything in your life that contributes to your personal well-being, rate your overall quality of life during the past seven days. (Outstanding, Very good, Good, Fair, Poor, Very poor)

Utility Assessment. Patients completed the UTiter II computer utility assessment. This is a computer program consisting of three utility assessment exercises that the subject completed with the assistance of the interviewer as needed. Utilities were measured with three direct techniques: the standard gamble, the time tradeoff, and a rating scale.

The standard gamble (SG) technique quantifies preferences by determining what chance of death the patient is willing to take to be freed from symptoms. The patient is asked to choose between continuing in his current state of health (with the associated symptoms) and an imaginary treatment that will completely cure the symptoms but carries a risk of death. The risk of death is altered until the patient is indifferent to the choice. For example, the patient may select the imagined treatment when there is a 10% chance of death but choose to continue in the current state of health if there is a 30% chance of death. The patient may be indifferent when the risk of death is 20%. The SG
utility is defined as 1.0 minus the risk of death at
the point of indifference, so that in this case, the
patient’s utility score is 0.80 (1.00–0.20).

The time tradeoff (TTO) metric quantifies
preferences by determining how many years a pa-
tient is willing to sacrifice to be freed from symp-
toms. The patient is asked to choose between
living a certain number of years in their current
state of health (with the associated symptoms) and
a lesser number of years but in perfect health. For
example, let us assume that an elderly patient
in poor health might reasonably expect to live 5
more years in his or her current state of health. He
or she might be willing to “trade” that amount of
time for 4 years of perfect health, but unwilling to
do so for just 2 years of perfect health. He may be
indifferent when he is offered 3 years of perfect
health. The TTO utility is defined as the number of
symptom-free years divided by the number of
years with symptoms (at the point of indifference).
In this example, the utility is 0.60 (3/5).

The rating scale (RS) technique is the simplest
to explain and is the least adherent to the axioms
of utility theory. We include its use because we
are only seeking to compare utility scores as an
outcome measure, rather than using them in a
formal decision analysis model. In its most ele-
mentary form, a visual analog scale is used, with
the worst possible outcome at one end and the
best possible outcome at the other. The patient
indicates the point on the scale that corresponds
to his or her state of health. The RS utility is
determined by dividing the actual distance from
the worst possible outcome to the point the
patient has indicated by the overall length of
the scale. For example, if the patient has
indicated that he or she feels his or her health
state is intermediate between the two ends of the
spectrum, the utility is 0.50.

### Table 1. Demographics.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Overall</th>
<th>Numerate</th>
<th>Nonnumerate*</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. patients</td>
<td>18</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>Age (mean)</td>
<td>59.6</td>
<td>63.8</td>
<td>57.1</td>
</tr>
<tr>
<td>Sex (% male)</td>
<td>88.2</td>
<td>100.0</td>
<td>75.0</td>
</tr>
<tr>
<td>Race (% white)</td>
<td>76.5</td>
<td>77.8</td>
<td>75.0</td>
</tr>
<tr>
<td>Medical comorbidity (% with ≥1)</td>
<td>47.0</td>
<td>44.4</td>
<td>55.6</td>
</tr>
<tr>
<td>Advanced stage (%)</td>
<td>81.3</td>
<td>75.0</td>
<td>87.5</td>
</tr>
<tr>
<td>Tumor site (% oral/oropharyngeal)</td>
<td>58.8</td>
<td>66.7</td>
<td>50.0</td>
</tr>
<tr>
<td>Global QOL (% “good”)</td>
<td>38.9</td>
<td>44.4</td>
<td>33.3</td>
</tr>
</tbody>
</table>

*No differences were statistically significant.

### Numeracy.

Numeracy was then assessed by use of three validated questions from Woloshin et al on a self-completed form. Subjects who answered

![Image of graphs showing utility scores for all patients and numerate vs. non-numerate patients.](image-url)
two or three questions correctly were classified as numerate. Subjects who answered no questions or only one question correctly were classified as innumerate. The questions are:

1. How many times would a coin flip come up heads if flipped 1000 times?
2. How much is one percent of 1000?
3. One is what percent of 1000?

**Analysis.** The sample was stratified by numeracy for analysis. Patients were additionally grouped
by responses to the single global QOL question as having “good” (outstanding, very good, or good) or “bad” (fair, poor, or very poor) overall QOL. The mean utility score of each group (“good”/“bad”) was compared between numerate and nonnumerate patients for each of the three utility assessment methods using Student’s t test. P values were two-sided and values of less than .05 were considered significant. One would expect patients in the “good” QOL group to have higher mean utility scores than patients in the “bad” QOL group. We hypothesized that for numerate patients, this would be true, but not for nonnumerate patients.

Pearson correlations between utility scores and observer-rated functioning were calculated and compared for the numerate and nonnumerate patients. Finally, correlations between utility scores using the three different measuring techniques were then calculated and compared between the numerate and nonnumerate patients.

Data were entered into a Microsoft Access database with double-entry verification techniques to ensure accuracy. The Human Subjects Review Committee at the University of Washington approved this study. All analyses were performed using STATA 6.

RESULTS

Eighteen patients with newly diagnosed head and neck cancer were enrolled in the study. The patient group was predominantly men and predominantly white and had a mean age of 59 years old (Table 1). There were more patients with advanced tumors. Fifty-nine percent of patients had oral or oropharyngeal primary tumors. Forty-seven percent of patients had one or more medical comorbidities. Half of the patients were classified as numerate.

---

Table 2. Spearman correlations between utility scores measured with three methods by numeracy.

<table>
<thead>
<tr>
<th></th>
<th>Numerate</th>
<th>Nonnumerate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rating scale</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard gamble</td>
<td>0.90</td>
<td>0.73</td>
</tr>
<tr>
<td>Time tradeoff</td>
<td>0.93</td>
<td>-0.33</td>
</tr>
<tr>
<td>Global QOL</td>
<td>0.80</td>
<td>0.23</td>
</tr>
<tr>
<td>Standard gamble</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time tradeoff</td>
<td>0.90</td>
<td>-0.62</td>
</tr>
<tr>
<td>Global QOL</td>
<td>0.58</td>
<td>0.56</td>
</tr>
<tr>
<td>Time Tradeoff</td>
<td>0.71</td>
<td>-0.55</td>
</tr>
</tbody>
</table>

*Moderate correlation = 0.2–0.4; strong correlation =>0.4.

---

Table 3. Spearman correlations of utility scores to clinician-rated function by numeracy.

<table>
<thead>
<tr>
<th></th>
<th>Numerate</th>
<th>Nonnumerate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinician rated function to</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rating scale</td>
<td>0.57</td>
<td>-0.16</td>
</tr>
<tr>
<td>Standard gamble</td>
<td>0.41</td>
<td>0.06</td>
</tr>
<tr>
<td>Time tradeoff</td>
<td>0.42</td>
<td>-0.13</td>
</tr>
<tr>
<td>Global QOL</td>
<td>0.58</td>
<td>0.26</td>
</tr>
</tbody>
</table>

On the basis of the responses to the global QOL question, 39% of the patients had good to outstanding overall QOL. The remaining 61% had fair to very poor QOL. There were no significant differences between numerate and nonnumerate groups for any of the demographic or disease variables.

When the sample was divided on the basis of responses to the global QOL question, there was no difference in mean utility scores for the “good” QOL and “bad” QOL groups by use of the TTO when both numerate and nonnumerate patients were combined for the analysis. When only numerate patients were included in the analysis, the utility scores of the group with “good” QOL were significantly higher than those of the group with “bad” QOL. This was the expected outcome. When only nonnumerate patients were analyzed, however, not only was there no significant difference between the “good” and “bad” QOL groups, but those with “bad” QOL actually had a higher mean utility score (Figure 1). Similar trends held to a lesser degree for the other measures of utility (Figures 2 and 3).

Correlations between scores on the three utility instruments were stronger for numerate than nonnumerate patients (Table 2). This indicates more consistency in the responses of the numerate patients. Similarly, utility scores of numerate patients correlated more strongly with observer-rated functioning (Table 3). Although observer-rated assessment of function is far from a “gold standard,” the strength of these correlations shows that utility scores have more face validity for numerate patients than nonnumerate ones.

DISCUSSION

We evaluated the effect that comfort with numerical expressions (numeracy) has on a patient’s ability to provide useful outcomes data on utility assessment instruments in a sample of head and neck cancer patients. We found that only half of
these patients were numerate and that numeracy strongly influenced the quality of utility-based outcomes. These data support the findings of Woloshin et al, who suggest that only a fraction of patients are comfortable with quantitative tasks and that innumerate patients’ responses to utility measures may be invalid.3

We elected to focus on utilities in this study because the existing literature suggested that the impact of numeracy on utility assessment was likely to be strong. Our findings confirmed our hypothesis. Although no “gold standard” for measuring QOL and evaluating health states is available, we showed better correlations between utility scores and objective measures of function for numerate patients. Furthermore, we showed that between three measures of utility that purported to measure the same thing, there was less consistency of responses by nonnumerate subjects. Finally, we showed that among the numerate population, utility measures were capable of distinguishing “good” from “bad” global QOL by use of the TTO. Among the nonnumerate patients, utility measures were not. When data from numerate and nonnumerate patients were combined, utility scores did not show a significant difference between patients with “good” and “bad” global QOL.

The concept that quantitative skills affect function is not new. Two large community-based studies documented poor functional health literacy and numeracy, particularly among older populations.12,13 The impact of numeracy on a patient’s ability to understand risk and benefit information has been well established.7 Woloshin et al14 took the next step by showing that lack of quantitative skills has a negative impact on the patient’s ability to value health states.

Utility assessment is not commonly used in studying the impact of head and neck cancer and its treatments on QOL. This is largely because, as a method, it is not terribly sensitive to treatment-related changes in QOL and because utility scores are not particularly meaningful to providers. Previous authors have summarized other concerns about utility assessment. These include the impact of patients’ comfort with risk, preferences for time, and cognitive ability to make judgments.14

Utility assessment is, however, broadly used in valuing health states for other medical ailments.15 Utility scores have the advantage of being generalizable beyond a specific disease entity and interpretable because norms exist from population-based studies.15 They are particularly useful for cost-utility assessment, where they provide a generic outcome measure allowing comparisons of value across disease entities (ie, comparing the value of the influenza vaccine to the value of screening mammography).

There are several studies evaluating the level of numeracy in various populations.3,7–9 The three-question numeracy instrument was commonly used. Previous studies found that only a small fraction of subjects answered all three questions correctly even in highly educated samples.3,9 Quantitative skill is, however, likely a continuum. Patients answering two of three correct are likely more comfortable with numerical skills than those answering zero or one correctly. Very few patients in this study answered all three questions correctly. The small sample size of this study did not allow for comparisons of those answering all three correctly with those answering one or none correctly. Comparison of the extreme ends of quantitative skill would allow for less misclassification than our dichotomous schema. Accordingly, the classification scheme used here might have impaired our ability to detect real differences between the two groups on the SG and RS as a result of misclassification of nonnumerate patients as numerate. In addition, the simple numeracy questionnaire is essentially a screening tool and may allow for some inherent misclassification despite its general validity. Nonetheless, our classification of numerate or nonnumerate provided a meaningful way to select out a population of patients for whom utility assessment is a poor measure of actual health status.

It should also be recognized that whereas the U-Titer II is a validated method of assessing utilities, lack of computer literacy may confound our data.5 The impact of computer literacy has not been studied in regard to these exercises, yet it is reasonable to assume that nonnumerate patients are less likely to be comfortable with computers. This in turn may have a further impact on the quality of responses by nonnumerate patients to the utility assessments used in this study.

The impact of numeracy on patient-based outcomes has broad ramifications. Elimination of nonnumerate patients from a study seems desirable to eliminate dubious data. The results of studies enrolling only numerate patients, however, may not be generalizable to a broader population. Prior studies have correlated advanced age and lack of education with poor health literacy and numeracy.12,13,16 There is concern that numeracy is, therefore, a surrogate measure of
socioeconomic status and education. There is likely some truth in this assertion. Other studies have correlated QOL with level of education. Combining these correlations implies that non-numerate patients may have lower QOL in general than do their numerate counterparts. Studies selectively excluding patients with a lower overall QOL would produce erroneously high valuations of health states.

The present finding that utility scores did not differ between patients with “good” and those with “bad” QOL when data from numerate and non-numerate patients were combined also holds broad implications for QOL assessment in patients with head and neck cancer. Many frequently used questionnaires require patients to rank responses on 5- or 7-point scales. This task may seem trivial, but our data showed negative correlations between utility scores and objective measures of function for some nonnumerate patients. This means that some patients may have reversed their responses (e.g., giving higher scores to worse health states). Woloshin et al. showed that nonnumerate subjects have a higher tendency to improperly order the severity of illness in hypothetical descriptions of health states. The implications of our findings and those of Woloshin is that even the ordering of responses on traditional psychometric questionnaires may be affected by lack of numeracy. This and the strong correlations between numeracy and literacy may have a negative impact on the quality of responses by nonnumerate patients to traditional QOL questionnaires. We did not specifically test that hypothesis in this study. Future research is needed to directly assess these relationships because this may help to explain why traditional QOL instruments have been unable to detect global QOL differences resulting from various treatments for head and neck cancer.

**CONCLUSIONS**

This study has shown that patients’ lack of facility with numerical expressions impairs their ability to provide meaningful data on QOL assessment performed with utility instruments. Furthermore, half of head and neck cancer patients in this study were not numerate. These results may help to explain the shortcomings of traditional QOL assessment in head and neck cancer.

**REFERENCES**