Quality of Life in Laryngeal Cancer Patients

By

G. Jolie Ringash

A thesis submitted in conformity with the requirements for the degree of Master of Science, Graduate Department of Community Health, University of Toronto

© Copyright by G. Jolie Ringash 1999
The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author’s permission.

L’auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L’auteur conserve la propriété du droit d’auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-45919-5
Abstract

ABSTRACT

Quality of Life in Laryngeal Cancer Patients
Master of Science 1999
G. Jolie Ringash
Department of Community Health, University of Toronto

Purpose:
To estimate the magnitude of difference in quality of life (QOL) that is noticeable to patients.

Materials and Methods:
Laryngeal cancer patients (n=120, 83% male, mean age=65) completed a health questionnaire, FACT-H&N, and two utility assessments, TTO and DATE. Paired participants subjectively rated their own QOL as compared to each other.

Results:
FACT-H&N was more highly correlated with Karnofsky (r=.43, p=.0001) than with the TTO (r=.29, p=.002) or DATE (r=.35, p=.0003). Subjective comparison ratings were positively skewed. The threshold for ratings of "a little bit better" rather than "just the same" was about 6 units for FACT-H&N, and about 0.05 units for either TTO or DATE. The thresholds for ratings of "a little bit worse" were significantly lower (-12, -0.06, -0.14 units respectively).

Conclusion:
An arithmetic QOL difference of 5 to 10% was noticeable. Participants were sensitive to a smaller positive than negative difference. Health questionnaire and utility measures may be perceived differently.
Acknowledgements

This thesis is dedicated to all those who encouraged me to ask questions, and to my husband Glen Bandiera for joining me in my quest for answers.

The author gratefully acknowledges permission to reproduce as appendices the FACT-H&N version 4 and the laryngeal cancer chapter of the TNM Classification of Malignant Tumours, Fifth Edition.
# TABLE OF CONTENTS

i. Face Sheet i

ii. Abstract ii

iii. Acknowledgements iii

iv. Table of Contents iv

v. List of Tables viii

vi. List of Figures ix

vii. List of Appendices x

1. Overview 1

#### 1.1 Purpose
- Minimal Important Difference for FACT-H&N
- Minimal Important Difference for Utility Measures
- Comparison of Health Questionnaires and Health Utilities
- Comparison of Two Versions of the Time Trade-off
- Predictors of QOL

#### 1.2 Relevance of MID 4
- Interpretation of QOL Data
- Design of Phase III Trials
- Decision Analysis
- Institutional Outcome

2. Background 7

#### 2.1 Laryngeal Carcinoma 7
- Incidence and Lethality
- Etiology
- Predictive Factors
- Importance of Quality of Life
- Medical Decision Making
- Tradeoffs Between Quality and Quantity of Life

#### 2.2 Patient Population 11
- Princess Margaret Hospital
- Larynx Cancer Treatment Protocols
- Follow-up Protocols

#### 2.3 Prior QOL Results in Laryngeal Cancer 13
- Health Questionnaire Results
- FACT-H&N Results
- Utility Results

#### 2.4 Health Questionnaires for Laryngeal Cancer 17
- Requirements for a Health Questionnaire in this Study
- Choice of Disease-Specific Health Questionnaire
- The FACT-H&N Questionnaire

#### 2.5 Health Utility Measurement 21
- Requirements for a Utility Measure in this Study
- Choice of Utility Instrument
- An Experimental Utility Technique
- Anchors and Time Interval
TABLE OF CONTENTS (CONTINUED)

2. Background (Continued)
   2.6 Estimation of the Minimal Important Difference
      A. Change over Time (Intra-Patient Method)
      B. Differences Between Individuals (Inter-Patient Method)
      C. Applications to Cancer Patients
      D. Choice of Method of Estimating MID

3. Methods
   3.1 Population
      A. Potential Participants
      B. Inclusion Criteria
      C. Exclusion Criteria
   3.2 Baseline Variables
      A. FACT-H&N Health Questionnaire
      B. Utility Measures
      C. Demographic, Tumour and Treatment Characteristics
      D. Performance Status
   3.3 Group Meetings
      A. FACT-H&N Health Questionnaire
      B. Utility Measures
      C. Inter-patient Comparisons
   3.4 Unanticipated Changes
      A. Judgemental Approach
      B. Statistical Approach
      C. Management of Outliers
   3.5 Statistical Analysis
      A. Sample Size
      B. Analysis of Minimal Important Difference
      C. Test of Robustness
      D. Relationship Between Measures of QOL
      E. Generalizability
   3.6 Ethics
      A. Informed Consent
      B. Impact on Patient Care
      C. Data Management and Reporting
      D. Ethics Committee Approval

4. Results
   4.1 Recruitment
      A. Degree of Participation
      B. Reasons for Patient Exclusions
      C. Reasons for Declining Participation
      D. Reasons for Withdrawals
      E. Representativeness of the Sample
   4.2 Baseline Results
      A. Qualitative Results
      B. Quantitative Results
      C. Correlation Between QOL Measures
      D. Clinical Predictors of QOL Scores
      E. Discrimination Between Known Groups
TABLE OF CONTENTS (CONTINUED)

4. Results (Continued)

4.3 Subjective Comparison Ratings
   A. Analyzable Data
   B. Distribution of Differences and Comparison Ratings
   C. Qualitative Observations
4.4 Estimates of Minimal Important Difference
   A. Correlation for Arithmetic Differences and Subjective Comparisons
   B. Asymmetry of MID
   C. "Positive" and "Negative" Estimates of MID
4.5 Tests of Robustness
   A. Subgroup Estimates of MID
   B. Relationship Between Item Number and MID
   C. Applicability of MID Estimates to Individuals

5. Discussion

5.1 Estimation of Minimal Important Difference
   A. Magnitude of the Estimates
   B. Asymmetry of Positive and Negative MID
   C. Validity and Precision of MID Estimates
5.2 Subjective Comparison Ratings
   A. Positive Skew
5.3 Utility Assessments
   A. The "Perfect Health" Anchor State
   B. Linearity of the Utility Function with Time
   C. Advantages and Disadvantages of the DATE
5.4 Health Questionnaires
   A. Institutional Outcome and Health Questionnaire Scores
   B. Test-Retest Reliability of the FACT-H&N
   C. Gender Differences in QOL
5.5 Relationship Between FACT-H&N and Health Utilities
   A. Can Health Questionnaire Scores Predict Utilities
5.6 Future Directions

6. Conclusion

7. Tables

Table 1: Characteristics of Participants and Non-participants
Table 2: Mean Health Questionnaire and Utility Scores
Table 3: Mean TTO and DATE Scores Excluding Patients Who Have or Refuse "Perfect Health"
Table 4: Reasons Stated for Declining Perfect Health (N = 18)
Table 5: Correlation Between QOL Measures
Table 6: Correlation Between Arithmetic Differences on FACT-H&N Subscales and their Corresponding Subjective Comparison Ratings
Table 7: Correlation Between Differences in QOL Measures
Table 8: Estimates of Positive and Negative MID
Table 9: Positive and Negative MID for Subscales of FACT-H&N
Table 10: Positive and Negative MID as Percentage of Total Instrument Score
### Table of Contents (Continued)

#### 8. Figures
- Figure 1: Plot of FACT-H&N #1 versus #2 for Visual Inspection of Change
- Figure 2: Distribution of Difference Scores for FACT-H&N
- Figure 3: Frequency Distribution for Subjective Comparison of Overall Well-being
- Figure 4: Comparison of Arithmetic Difference in FACT-H&N versus Subjective Comparison of Overall Well-being
- Figure 5: Comparison of Arithmetic Difference in Time Trade-off versus Subjective Comparison of Overall Well-being
- Figure 6: Mean Difference in FACT-H&N for Each Level of Subjective Comparison
- Figure 7: Mean Difference in TTO for Each Level of Subjective Comparison
- Figure 8: Mean Difference in DATE for Each Level of Subjective Comparison
- Figure 9: Mean Difference on the H&N Subscale of FACT-H&N for Each Level of Subjective H&N Comparison
- Figure 10: Estimates of Positive MID for FACT-H&N from Subgroup Analyses
- Figure 11: Estimates of Negative MID for FACT-H&N from Subgroup Analyses
- Figure 12: Cumulative Probability of a Patient Rating Himself at Least “A Little Bit Better” for a Given Arithmetic Difference in TTO

#### 9. References

#### 10. Appendices
- A. UICC Staging Criteria for Laryngeal Carcinoma
- B. Introductory Letter to Prospective Patients
- C. Patient Consent Form
- D. FACT-H&N Version 4
- E. Time Trade-off (TTO) Semi-structured Interview Script
- F. Daily Active Time Exchange (DATE) Interview Script
- G. Karnofsky Performance Status Scale
- H. Subjective Comparison Rating Form
- I. Princess Margaret Hospital Clinical Trials Committee Approval
- J. University of Toronto Oncology Human Review Committee
List of Tables

Table 1: Characteristics of Participants and Non-participants

Table 2: Mean Health Questionnaire and Utility Scores

Table 3: Mean TTO and DATE Scores Excluding Patients Who Have or Refuse “Perfect Health”

Table 4: Reasons Stated for Declining Perfect Health (N = 18)

Table 5: Correlation Between QOL Measures

Table 6: Correlation Between Arithmetic Differences on FACT-H&N Subscales and their Corresponding Subjective Comparison Ratings

Table 7: Correlation Between Differences in QOL Measures

Table 8: Estimates of Positive and Negative MID

Table 9: Positive and Negative MID for Subscales of FACT-H&N

Table 10: Positive and Negative MID as Percentage of Total Instrument Score
List of Figures

Figure 1: Plot of FACT-H&N #1 versus #2 for Visual Inspection of Change

Figure 2: Distribution of Difference Scores for FACT-H&N

Figure 3: Frequency Distribution for Subjective Comparison of Overall Well-being

Figure 4: Comparison of Arithmetic Difference in FACT-H&N versus Subjective Comparison of Overall Well-being

Figure 5: Comparison of Arithmetic Difference in Time Trade-off versus Subjective Comparison of Overall Well-being

Figure 6: Mean Difference in FACT-H&N for Each Level of Subjective Comparison

Figure 7: Mean Difference in TTO for Each Level of Subjective Comparison

Figure 8: Mean Difference in DATE for Each Level of Subjective Comparison

Figure 9: Mean Difference on the H&N Subscale of FACT-H&N for Each Level of Subjective H&N Comparison

Figure 10: Estimates of Positive MID for FACT-H&N from Subgroup Analyses

Figure 11: Estimates of Negative MID for FACT-H&N from Subgroup Analyses

Figure 12: Cumulative Probability of a Patient Rating Himself at Least “A Little Bit Better” for a Given Arithmetic Difference in TTO
List of Appendices

A. UICC Staging Criteria for Laryngeal Carcinoma

B. Introductory Letter to Prospective Patients

C. Patient Consent Form

D. FACT-H&N Version 4

E. Time Trade-off (TTO) Semi-structured Interview Script

F. Daily Active Time Exchange (DATE) Interview Script

G. Karnofsky Performance Status Scale

H. Subjective Comparison Rating Form

I. Princess Margaret Hospital Clinical Trials Committee Approval

J. University of Toronto Oncology Human Review Committee
1. Overview

1. Overview

1.1 Purpose

The primary objective of this project was to estimate the magnitude of the minimal important difference for some measures of treatment outcome in laryngeal cancer: a disease-specific health-related quality of life (QOL) questionnaire, an established health utility instrument, and an experimental measure of health utility. In accordance with previous investigators, the minimal important difference (MID) was defined as, “the smallest difference in measured health status that signifies an important rather than a trivial difference in patient symptoms.”

Secondary objectives included a comparison of two versions of the time trade-off method of measuring health utilities; an assessment of the correlation between the health questionnaire, utility and performance status instruments; and a regression analysis to identify patient, disease and treatment factors which predicted QOL.

A. Minimal Important Difference for FACT-H&N

Following literature review (section 2.4.B), the Functional Assessment of Cancer Therapy Head and Neck questionnaire version 4 (FACT-H&N) was chosen as a psychometrically validated disease-specific QOL instrument relevant to the clinical domain. The Functional Assessment of Chronic Illness Therapy (FACIT) measurement system, developed by Cella, is a modular questionnaire system consisting of a general component, FACT-G, with additional modules for a variety of chronic illnesses.

The FACT-H&N is a multi-dimensional index which provides a total score by summing separate subscores for physical, emotional, social, functional, and head and neck cancer
dimensions (section 2.4.C). In this research project, the MID was estimated for the measure as a whole (FACT-H&N), for the general portion of the measure (FACT-G), and for each of the five subscales individually.

B. Minimal Important Difference for Utility Measures

Utility measures are intended to quantify not only health, but also its value to the individual. We selected the time trade-off (TTO) as our established utility measurement technique. The TTO is a holistic measure of utility that is expressed by the patient in a personal interview. The method requires the patient to consider a given period of time in his current health state, and then to consider what period of time in “perfect health” would be of equal value to him. The World Health Organization (WHO) defines “perfect health” as, “a state of complete physical, mental and social well-being, not merely the absence of disease and infirmity.” The patient gives up time in order to gain quality of life (section 2.5.B). In this study, the MID was also estimated for an experimental method of utility assessment, the Daily Active Time Exchange (DATE), which may have advantages for use in cancer patients (see section 2.5.C).

C. Comparison of Health Questionnaires and Health Utilities

The World Health Organization (WHO) defines QOL as,

“an individual’s perception of their position in life in the context of the culture and value systems in which they live and in relation to their goals, standards and concerns. It is a broad ranging concept affected in a complex way by the person’s physical health, psychological state, level of independence, social relationships, and their relationships to salient features of their environment.”

Health-related QOL is defined as,

“optimum levels of mental, physical, role (e.g. work, parent, career, etc.) and
social functioning, including relationships, and perceptions of health, fitness, life satisfaction and well-being. It should also include some assessment of the patient's level of satisfaction with treatment outcome and health status and with future prospects. It is distinct from quality of life as a whole, which would also include adequacy of housing, income and perceptions of immediate environment.6

QOL can be measured with patient-centred, decomposed, multi-dimensional descriptive health questionnaires, most of which have been developed by clinicians or psychometricians.

Health utilities are defined as "numbers that represent the strength of an individual's preference for particular outcomes when faced with uncertainty."7 Expected utility theory originated in the field of economics8 and has been applied to health mainly by health policy makers. Utility measurements are most commonly global and unidimensional. They may sometimes be elicited from substitute decision makers, such as taxpayers, rather than from patients.

Despite these differences, both health questionnaires and utilities provide an index of the "goodness" of life. Direct comparisons of the two concepts have demonstrated that the relationship between them is not simple.9,10 Not all authors consider utility instruments as a subset of QOL measurement techniques, and the issue of whether health questionnaire scores can be converted to utilities remains unsettled. In laryngeal cancer, a direct comparison has not been made. As a secondary outcome, we wished to determine to what degree health utilities measured by the TTO and DATE would correlate with FACT-H&N scores, and with performance status as measured by the Karnofsky scale.11

D. Comparison of Two Utility Assessment Techniques

Considerable methodological questions persist in the field of utility research, and currently
available measurement techniques are not ideal. An alternative daily formulation of the TTO, in which patients are asked to give up a portion of their time each and every day for the rest of their lives in order to improve their health, was proposed by Buckingham\textsuperscript{12} as a potential improvement. In a comparison with the TTO, this method showed a higher response rate, a greater willingness for more individuals to trade at least some time, and better construct validity. We chose to directly compare this method, which we have termed the Daily Active Time Exchange (DATE), with the standard TTO in laryngeal cancer patients. Correlation was used as a measure of agreement. In addition, construct validity was assessed by using regression techniques to determine whether patient, tumour, or treatment factors predicted TTO and DATE scores in the expected manner. Finally, the responsiveness of DATE to small differences between individuals was compared to that of the TTO by estimating the MID for both methods.

E. Predictors of QOL

Regression analysis can identify patient, disease and treatment factors that are significantly associated with health questionnaire, utility or functional status scores. Such factors could be used at the time of diagnosis to help identify individuals at risk for negative outcomes and allow for psychosocial interventions or alternative treatment strategies. In addition, if factors expected to predict for low scores are confirmed, this information supports the construct validity for the instruments.

1.2 Relevance of MID

A. Interpretation of QOL Data

Few studies of QOL have been completed for laryngeal cancer patients. Indeed, no published phase III randomized trials of different approaches to larynx cancer therapy have yet measured
1. Overview

QOL endpoints prospectively. One ongoing three arm Intergroup trial for patients with advanced larynx cancer is comparing radiation alone, concomitant radiation plus chemotherapy, or a strategy of induction chemotherapy followed by either definitive radiotherapy for responders or surgery for non-responders. This trial is measuring QOL on the FACT-H&N prospectively, and when completed may show changes in QOL over time within the three treatment arms, as well as document long-term QOL for each of the three strategies. In order to interpret the results of this and other trials, however, it may be helpful to have empiric data regarding what difference in QOL scores on FACT is clinically significant and noticeable to patients; that is, to have an estimate of MID.

B. Design of Phase III Trials

The oncology community recognizes that QOL is a major factor in treatment decision-making for laryngeal cancer. For early disease, options may include conventional radiotherapy (RT), hyperfractionated RT, or partial laryngectomy. In more advanced disease, total laryngectomy with or without voice reconstitution, radiation alone with conventional or altered fractionation, combined chemoradiation, or induction chemotherapy followed by local RT or surgery are all options. Two large, randomized trials are presently ongoing but will not answer all relevant questions regarding therapy. Future trials are likely, and QOL endpoints should be included as secondary or even primary endpoints. In order to plan the sample size of such trials, it will be necessary to have estimates of MID for the outcomes.

C. Decision Analysis

One attempt has already been made to answer the question of best treatment approach for larynx cancer using decision analysis, but it was hampered by the lack of good data on health utilities in these patients. Combined with the additional information regarding survival and local control
outcomes which should be forthcoming from current trials, accurate utility measurements in a large number of larynx cancer patients could be used to update this initial analysis and to support future planning. In particular, decision analysis could be very valuable in comparing primary ablative (surgical) versus conservative (radiation, with or without chemotherapy) treatment, since previous attempts to implement a randomized trial comparing these different approaches have failed (Dr. Libni Eapen, Ottawa Regional Cancer Centre, Ottawa, Canada. Personal Communication. June 28, 1999). In such an analysis, it would be vital to have accurate estimates of chronic health. Our study measured health utilities in a large number of patients of different stages who were treated with primary radiotherapy. Additional such work in a centre that frequently employs primary surgical management would complement our results and allow for a more data-based decision analysis.

D. Institutional Outcome

Finally, we wished to measure the well-being of a large sample of patients treated for laryngeal cancer at a single tertiary referral centre. Previous studies at our institution have examined treatment outcomes for laryngeal cancer patients in terms of overall survival and local control16, but have not measured QOL. In addition, the use of regression analysis might identify groups of patients who have an inferior QOL outcome from treatment, and for whom alternative treatments or additional supportive care might be worth considering. Moreover, any study of QOL helps to raise awareness of this aspect of medical care and can indirectly result in a more humane health care system.
2. BACKGROUND

2.1 Laryngeal Carcinoma

A. Incidence and Lethality

The larynx is the most common site of malignancy in the head and neck region, and accounts for approximately 1% of all incident cancers in Canada. In 1997, an estimated 1,290 new cases were diagnosed, and 530 deaths occurred as a result of the disease,\textsuperscript{17} for an overall case fatality rate of 0.41. The peak incidence is in the sixth decade.\textsuperscript{18} The tumour occurs four times more frequently in males than in females,\textsuperscript{17} with an age-standardized incidence rate of 7/100,000 in males.\textsuperscript{17} Although the incidence rate in males decreased by an average of 1.6% annually between 1985-1992,\textsuperscript{17} no significant change in mortality rate has been observed.

B. Etiology

Tobacco smoking is the major cause of laryngeal cancer, with cigarette smoking accounting for between 82\%\textsuperscript{19} to 95\%\textsuperscript{20} of reported cases. A dose-response relationship has been observed.\textsuperscript{21} A similar risk applies to the use of pipes or cigars.\textsuperscript{22} Alcohol augments the risk, which is 75\% higher in users of both substances compared to those using only one.\textsuperscript{23} The supraglottic subsite is over-represented in heavy alcohol users.\textsuperscript{24} Heavy marijuana smoking also seems to be an etiological factor in some patients.\textsuperscript{25} Other possible contributors include vocal abuse; gastroesophageal reflux; exposure to ionizing radiation; or extended inhalation of wood dust, asbestos or nitrogen mustard.\textsuperscript{26}
2. Background

C. Predictive Factors

Stage is the most important predictive factor for treatment outcomes in larynx cancer. Tumour category (location and size) predicts for the likelihood of local recurrence after treatment, and overall UICC (International Union Against Cancer)\textsuperscript{27} stage, which incorporates both primary tumour and nodal characteristics [Appendix A], predicts locoregional control and survival.\textsuperscript{25} Pathologic grade plays a lesser role. Overall, women tend to have a more favourable prognosis than do men.\textsuperscript{16} For patients treated with radiotherapy, smoking cessation may improve local control.\textsuperscript{28}

D. Importance of QOL

Considerable interest has been given to QOL issues for laryngeal cancer patients. The disease is relatively unique, in that it directly threatens the vocal apparatus and can lead to communication barriers. Early studies of morbidity and well-being focused on vocal rehabilitation and voice quality.\textsuperscript{29, 30} However, evidence has since shown that differences in voice handicap explain only a small part of the differences in QOL between patients.\textsuperscript{31} For patients who have undergone laryngectomies, the presence of a stoma can also have significant effects,\textsuperscript{32} perhaps because it is socially obvious and can interfere with swimming, showering, blowing the nose, and sexual activity.\textsuperscript{33}

In an interview study, the greatest number of laryngectomized patients indicated the following issues as being of importance to their QOL: physical consequences, interference with social activities, communication impairment, lifestyle changes and functional status.\textsuperscript{34} Among patients treated with radiation, tumour category predicted QOL,\textsuperscript{35} and fractionation schedule did not seem to do so.\textsuperscript{36, 37} Whether QOL depends on the choice of radiation or surgery as primary treatment is unclear, with studies showing conflicting results.\textsuperscript{31, 38, 39} Contrary to expectations, improvement of
QOL with time since treatment has not been consistently shown.\textsuperscript{31, 38, 36}

E. Medical Decision Making

Multicentre, randomized controlled trials are a relatively recent arrival in larynx cancer, but are currently underway to better determine the ideal treatment strategy with regard to survival and local control. At present, most evidence is from single institution non-randomized series. Although certain treatment approaches may have advantages as discussed below, it is possible that ultimate survival is independent of the primary treatment modality employed. Even less is known about the QOL outcomes of different treatment strategies.

Currently, treatment decisions seem dependent on the physician’s specialty and geographic location. In a multi-national survey,\textsuperscript{40} otolaryngologists were more likely than radiation oncologists to recommend surgical management, including partial laryngectomy for early disease and total laryngectomy for advanced tumours. Specialists in Canada, Scandinavia and the United Kingdom recommended radiotherapy more frequently for both early and advanced tumours than did their counterparts in the United States and Australasia.

Early stage laryngeal cancer (UICC T1-2N0M0), which is localized to the larynx without fixation of the vocal cord, is by far the most common presentation. It is generally treated with radiotherapy alone, and has an excellent 5 year local control of 68-91\%.\textsuperscript{16} Conservative surgery can produce comparable local control and survival\textsuperscript{41} but results in inferior voice quality.\textsuperscript{42} For supraglottic tumours, conservative surgery may be used in preference to radiotherapy in some centres, particularly for bulky lesions,\textsuperscript{25} since an inverse relationship between tumour size and local control with radiation has been shown.\textsuperscript{43}

For T3 cancers, which cause fixation of the vocal cord, treatment is highly controversial. Five
year survival ranges from 55-72% for surgical treatment, which usually requires total laryngectomy. A small, highly selected subset of patients with disease limited to the vocalis muscle may be safely treated with vertical hemilaryngectomy. Local control with conventional radiotherapy is 36-45% in the glottis and 55-82% in the supraglottis. Twice daily radiation may improve primary glottic local control rates to 67-71%. With close follow-up, surgical salvage provides overall local control of 57% in the glottis and up to 83% in the supraglottis. Survival ranges from 55% for glottic disease to 37% for supraglottic tumours.

The same issues exist for T4 disease, which extends beyond the larynx. Local control and survival rates tend to be lower. With radiation alone, Harwood found local control of 56% and, with surgical salvage, survival was 49%. Total laryngectomy provides survival rates between 25-54%. One retrospective study of T3 and T4 laryngeal cancer patients showed significantly better local control and survival for the T4 subgroup with laryngectomy as compared to radiation therapy.

As an alternative to radiation alone for advanced tumours (T3-4N0-3M0), preservation of the larynx may be achieved by utilizing a strategy of induction chemotherapy, with responders going on to definitive radiotherapy and non-responders receiving surgery. In a randomized trial involving a group of advanced glottic and supraglottic patients, this strategy produced a 53% overall survival, which did not differ significantly from 56% with total laryngectomy. After 2 years of follow-up, 64% of patients on the induction chemotherapy arm retained their larynx. However, it is not yet known whether the induction chemotherapy plus radiation approach provides any advantage over radiotherapy alone. An ongoing randomized trial is comparing this sequential strategy to radiotherapy alone or concurrent chemoradiotherapy.

Management of the neck nodes is also an important part of treatment for larynx cancer. Lymph node involvement is uncommon for early glottic tumours but may be present with T1-2
supraglottic disease. Locally advanced tumours have a higher rate of neck node disease. In the clinically involved neck, or where the risk of subclinical involvement is known to be high, neck dissection is added to the surgical approach. Where primary radiation is used, the lymphatics at risk can be included in the treatment fields. Planned neck dissection may be added to a primary radiotherapeutic approach for large (category N2b-3) clinically positive nodes.\textsuperscript{57} Post-laryngectomy radiotherapy may also be employed for positive margins, multiple positive nodes or extracapsular extension.\textsuperscript{25} Thus, in reality, series reporting outcomes with a given primary treatment approach have often used multiple modalities, the relative contributions of which are unknown.

F. Tradeoffs Between Quality and Quantity of Life

The primary decision in the management of advanced (T3 and T4) laryngeal cancer is between an ablative versus a conservative approach. Ablative surgery produces the highest rates of local control, and may provide a survival advantage. In contrast, the natural larynx can be conserved by utilizing radiation with or without induction chemotherapy. Surgery is then reserved for recurrence, and a larger proportion of patients retains their natural anatomy. As ongoing trials better delineate the survival outcomes of alternative strategies, a better understanding of their QOL outcomes will be necessary to make explicit the inherent tradeoffs between quality and quantity of life.

2.2 Patient Population

A. Princess Margaret Hospital

Princess Margaret Hospital, a member institution of Toronto’s University Health Network, is one of two tertiary/quaternary cancer care centres serving a referral base of 5 million people.
Approximately 7,000 new patients are seen at Princess Margaret each year, with about 300,000 total annual patient visits. The multidisciplinary head and neck oncology group comprises sixteen physicians (five radiation oncologists, seven surgical oncologists, two medical oncologists, one plastic surgeon, and a psychiatrist), two dentists, several nurses, a speech pathologist, a nutritionist and a social worker. Annually, approximately 250 new larynx cancer patients are seen and 2,000 follow-up visits occur (Dr. Bernard Cummings, Chief, Department of Radiation Oncology, Princess Margaret Hospital/University Health Network. Personal Communication. June 28, 1999).

B. Larynx Cancer Treatment Protocols

At Princess Margaret Hospital, early laryngeal cancer patients are staged with direct laryngoscopy and biopsy, indirect laryngoscopy and chest x-ray. For locally advanced or clinically node positive patients, CT and occasionally MRI scans are added. In the absence of pertinent systemic symptoms, bone scan and CT assessment of the liver and brain are not necessary.

Princess Margaret Hospital has a long tradition of laryngeal preservation. Primary radiotherapy is used for all early laryngeal cancers, and is offered to most patients with advanced tumours. However, primary surgery is considered for patients who prefer that strategy, or who are unable or unwilling to undergo regular follow-up. Patients with large clinically positive neck nodes (N2b-N3) are treated with combined therapy, involving initial radiotherapy to the primary tumour and lymph nodes, followed by modified radical lymph node dissection. Outside of clinical trials exploring altered fractionation, once daily radiation treatment is used. The standard dose for T1 and small T2 tumours is 51 Gy in 20 fractions. Larger primaries or patients with known or suspected involvement of neck nodes are treated to a dose of 70 Gy in 35 fractions. Induction or concurrent chemotherapy is not presently used routinely. At their initial visit, all
patients are strongly urged to forgo smoking.

C. Follow-up Protocols

The hospital policy for follow-up of laryngeal cancer patients requires indirect laryngoscopy monthly for the first year post-treatment, then every two months in year two. During year three, patients are examined every three to four months. Exams are then performed every six months in years four and five post-radiation, and subsequently annually. These regular follow-ups provide a convenient opportunity to contact patients and conduct outcome studies.

2.3 Prior QOL Results in Laryngeal Cancer

A. Health Questionnaire Results

Literature review revealed 10 studies that measured QOL exclusively in laryngeal cancer patients. Using the Psychosocial Adjustment to Illness Scale (PAIS) and a surgery-specific questionnaire for laryngectomized patients, Desanto found better psychosocial adjustment among 23 patients with partial laryngectomies as compared to 149 with total or near-total laryngectomies. In India, using a questionnaire in the local language, Deshmane observed severe QOL problems. Over 80% of the 50 laryngectomized patients had restricted their social activities, 78% of patients faced major financial difficulties due to inability to work, and 32% were able to communicate only by writing. Stewart assessed 80 patients with a general QOL instrument, the Medical Outcomes Study Short-Form 36 (SF-36), and found no significant differences in QOL between patients treated with radiation or surgery. He also failed to observe a relationship between QOL and time since treatment. In 46 patients, Morton used the Quality of Life Instrument for Head and Neck (QL-H&N) prospectively and found improvements in psychological distress and weight with time since treatment. He noted more difficulty in
2. Background

speaking among surgical patients than among those treated with radiation, but emotional well-being and psychological distress did not differ between the two groups. Terrell\textsuperscript{39} assessed survivors of a randomized laryngeal preservation trial\textsuperscript{39} using the SF-36 and the University of Michigan Head and Neck Quality of Life Questionnaire (HNQOL) and found less pain and better mental health among patients on the induction chemotherapy and radiation arm, as compared to those treated surgically. With the European Organization for Research into the Treatment of Cancer Quality of Life Questionnaire for Head and Neck Cancer (EORTC QLQ-C30) and an earlier version of its head and neck module, the H&N37, Hammerlid\textsuperscript{36} prospectively assessed 57 patient undergoing radiotherapy and found most QOL dimensions to have reached a nadir 1 month following treatment, and to have improved to above baseline levels by 1 year post-therapy. No differences were observed between the group with small tumours treated once daily and that with large tumours treated twice daily. Finizia\textsuperscript{60} observed no differences on the EORTC QLQ-C30/H&N-35 between a matched group of 14 successfully irradiated patients and 14 patients using tracheoesophageal speech after salvage laryngectomy for radiation failure. De Graeff\textsuperscript{61} prospectively studied 65 patients with Tis to T3 laryngeal cancer treated with radiotherapy alone. He found that fatigue, physical and emotional functioning, swallowing, social eating, speech, taste, dry mouth and sticky saliva all caused more problems six months after treatment than at diagnosis. By one year post-treatment, most scales had returned to above the baseline level, but dry mouth, sticky saliva and poor taste continued to be below the baseline level. Despite these subscale changes, no significant differences were observed in mean global QOL scores. On the EORTC QLQ-C30, no differences were observed at any time between patients with T1 versus T2 tumours. However, the H&N35 did show more problems on most scales at one year post-treatment as compared with baseline.

B. FACT-H&N Results

Previous studies of QOL have used earlier versions of FACT-H&N with slightly different total
2. Background

scores. Only one study using FACT-H&N included exclusively laryngeal cancer patients. In this small prospective study, QOL was assessed for newly diagnosed patients on four occasions from diagnosis until six months post-treatment. Groups of seven patients each underwent total laryngectomy, hemilaryngectomy or radiotherapy. Irradiated patients had earlier disease, and the total laryngectomy group had the most advanced tumours. A trend was found for better QOL in the irradiated group, but statistically significant differences between groups were not observed. On an earlier version of FACT-H&N, the mean score was 110/148 (74%) at 6 months post-treatment, with means of 87/112 (78%) on FACT-G and 22/36 (61%) on the head and neck module.

Three other studies have included laryngeal cancer patients with other head and neck cancer patients. List included 49 such patients in a cross-sectional study of 151 patients, and found overall means of 104/148 (70%) for FACT-H&N, 85/112 (76%) for FACT-G and 19/36 (53%) for the H&N subscale. In a study of 50 patients of whom 22 had laryngeal cancer, means were 116/156 (74%) for FACT-H&N, 88/112 (79%) for FACT-G and 28/44 (64%) for the H&N module. In the same patients, high H&N module scores were significantly associated with a diagnosis of larynx cancer as compared to other head and neck sites, and also with low ratings of dysfunction on the Performance Status Scale for Head and Neck Cancer (PSS-H&N). Patients who had low disfigurement ratings on PSS-H&N or who were married and did not live alone had higher FACT-G scores. In a study of 47 patients (of whom 15 had laryngeal cancer) who had undergone concomitant chemoradiotherapy for stage II-IV tumours, the mean FACT-H&N score was 117/164 (71%), with means of 89/120 (74%) for FACT-G and 28/40 (70%) for the H&N module.

C. Utility Results

Most previous studies had obtained utilities for larynx cancer scenarios from surrogates or had
rated voice quality. The first attempt to measure utility in laryngeal cancer patients was by Llewellyn-Thomas. She included 29 such patients in a study of 61 cancer patients. At the start of their treatment, participants used the standard gamble (SG) and rating scale (RS) to rate standardized, personal scenarios meant to represent their chronic health state after radiotherapy. This process was repeated when they actually reached the end of treatment. She observed significant changes over time, and for the whole group, the mean SG utility was 89 at the start, and 84 at the end of treatment. For the RS, the corresponding scores were 76 and 70. In a later study she asked 66 larynx cancer patients to rate standardized scenarios (which were intended to represent the chronic, post-radiotherapy health state) using the TTO and RS methods at both the start and end of therapy. No significant differences over time were observed with either method. Group mean TTO scores at the end of therapy for the “mild”, “moderate” and “severe” scenarios were 0.76, 0.64, and 0.40 with corresponding RS scores of 0.76, 0.57, and 0.24.

Only one previous study has asked larynx cancer patients to rate the utility of their personal health state. In this study, 10 previous patients indicated a mean utility of 0.61 using the RS method. They also assessed standardized T3 larynx cancer treatment scenarios as part of a trial of 39 individuals, most of them surrogates. For the scenario describing surgery the patients expressed utilities of 0.62, 0.65 and 0.45 by the SG, TTO and RS methods. Utilities for the radiation scenario were 0.61, 0.70 and 0.66 respectively. Clinicians and members of the general population generated higher values for the same scenarios.

To date, personal utilities have not been measured in a large sample of larynx cancer patients. To our knowledge, no other such trials are presently underway.
2. Background

2.4 Health Questionnaires for Laryngeal Cancer

A. Requirements for a Health Questionnaire in this Study

In this study, the priority was to measure long-term QOL for laryngeal cancer patients surviving their disease and therapy. An instrument was required which assessed issues of concern in the chronic phase of the disease, rather than acute changes during therapy. Since the majority of participants were expected to have undergone radiotherapy, it was especially important that the questionnaire deal adequately with known chronic radiation-induced toxicities. A modular instrument with a general portion relevant across cancer diagnoses would be beneficial, but was not an absolute requirement for the goals of this study. The primary role of QOL measurement was discriminative, measuring differences between laryngeal cancer survivors. Since the primary goal of the study was to estimate MID, and such an estimate of MID might later be used to interpret change within an individual over time, an instrument with good evaluative properties was also to be desired.

Because of the discussions between patients required in order to estimate MID, only English speaking patients were included in the study. Therefore, the QOL instrument needed to be available in English only, but needed to be culturally appropriate for the diverse Canadian population. For practical reasons, and also to avoid bias, a brief, self-assessed instrument was preferred.

The chosen instrument needed to have demonstrated reliability, validity and responsiveness to change. Finally, it was considered most practical to estimate MID for an instrument which is currently being used in larynx cancer trials, and which is likely to be widely used in future trials.

Controversy exists in the QOL field regarding the validity of a summary score for health
questionnaires. Psychometric theory suggests that it is not valid to summate across different domains of QOL, since the relative weightings of each domain are unknown, and may indeed vary between individuals. However, clinimetric scales designed for QOL and other related outcomes have traditionally summated across questions addressing a variety of issues, and have shown validity in their subsequent trials. Because of the nature of MID, it was judged that an index utilizing a summary score would be preferable, as it would provide greater simplicity in the interpretation of differences between individuals or over time, not only in this study but for future users of our data. Given the potential instruments available, that which best satisfied these requirements was chosen.

B. Choice of Disease-specific Health Questionnaire

Literature review identified eight multi-dimensional, head and neck specific health questionnaires (Ringash J. Assessment of quality of life instruments for head and neck cancer patients, 1999, manuscript in preparation). Critical appraisal of each questionnaire was carried out, considering issues of item generation, item reduction, sensibility, administration, scoring, reliability and validity.

Two instruments, the Quality of Life Instrument for Head and Neck Cancer (QL-H&N) and the Quality of Life Questionnaire for Advanced Head and Neck Cancer (QLQ) were eliminated from consideration due to inadequate information on development and sensibility.

A third instrument, the European Organization for Research into the Treatment of Cancer Quality of Life Questionnaire for Head and Neck Cancer (EORTC QLQ-C30/H&N-35) was eliminated for lack of published reliability and validity data. Although its general questionnaire, the EORTC QLQ-C30, is well-validated, the head and neck cancer specific module H&N35 has undergone substantial modifications, and at the time our trial was planned, the
current version had only been used in a single descriptive study. The results of three additional studies have since been published. The questionnaire is being used in one ongoing multicentre randomized trial. At the time we were choosing a questionnaire, it was felt that this otherwise promising instrument had not yet been adequately assessed for reliability and validity in a head and neck cancer population.

Fourth and fifth instruments, which both focus on radiotherapy issues, were initially appealing but lacked validation in the chronic, post-treatment situation. The Head and Neck Radiotherapy Questionnaire (HNRQ) and the Quality of Life - Radiation Therapy Instrument Head & Neck Module (QOL-RTI/H&N) would be worthwhile instruments, however, in a trial assessing QOL during the acute treatment phase.

A sixth instrument, the University of Michigan Head and Neck Quality of Life (HNQOL) instrument has also been modified, and limited data is available on the current version. Despite impressive early validation results, practical considerations also argued against the choice of this instrument, which is interviewer-administered, does not provide a single summary score, and requires the simultaneous use of a general QOL instrument since it does not cover all QOL domains.

A seventh instrument, the University of Washington Quality of Life Questionnaire (UW QOL) is currently being used in two multicentre randomized trials involving larynx cancer patients. This instrument is popular in North America, and serious consideration was given to its use. However, an assessment of the sensibility of the instrument suggested it would be most appropriately applied to patients who had undergone surgery, and that issues of potential importance to radiated patients were not as well covered. Additionally, it does not cover all dimensions of QOL and should be administered with a general QOL instrument.
The eighth and final instrument, the Functional Assessment of Cancer Therapy - Head and Neck (FACT-H&N) offered the advantages of modular design, self-administered format, a single summary score, and validation in the post-treatment phase. It was developed in the United States and its validation studies have all involved North American patients. It is currently being used for larynx cancer patients in a major multicentre randomized trial. Given these considerations, the FACT-H&N was chosen as the disease-specific health questionnaire for this study.

C. The FACT-H&N Health Questionnaire

FACT-G consists of 27 questions in four domains - physical (7), social/family (7), emotional (6), and functional (7). It is supplemented by a head and neck cancer specific subscale consisting of 11 questions, to make up the 38 item FACT-H&N. Two of the head and neck subscale items dealing with alcohol and tobacco use are not currently scored. The instrument specifies a time period of the past 7 days. Each response is rated from 0-4 on a Likert scale, with 0 described as “not at all” and 4 as “very much.” Scores are calculated separately for each domain, and an unweighted summary score is calculated for the FACT-G and the total FACT-H&N. The maximum score of 144 reflects the best possible quality of life.

The population used to develop the FACT-G focused initially on outpatients with common cancers (breast, lung and colon). However, it has been validated using in- and out-patients, as well as post-treatment patients, with mixed cancer diagnoses. The head and neck module was developed and validated on a mixed group of head and neck cancer patients (Dr. David Cella. Personal communication. 26 Feb. 1998). Item reduction used psychometric principles, including item response theory and factor analysis.

The head and neck cancer subscale was first published in 1996, and has been used to date in
eight published trials\textsuperscript{38, 62, 63, 64, 65, 66, 95, 96} and several abstracts. The FACT-H&N is currently being used in a major ongoing multicentre trial for laryngeal cancer in the USA and Canada.\textsuperscript{13}

The questionnaire format is easy to read, with instructions provided and consistent response options. Items are phrased as statements, each clear in meaning and consisting of less than 25 words, for example, “I am satisfied with how I am coping with my illness.” The completion time is about 5 minutes. Standardized scoring sheets are available.

Reliability has been assessed by both internal consistency and test-retest methods to be adequate to excellent for all subscales of FACT-G and by internal consistency for the head and neck subscale. Assessments of concurrent validity have used the Karnofsky score, the PSS-H&N,\textsuperscript{62} the UW-QOL,\textsuperscript{63, 65, 66} the HNRQ,\textsuperscript{66} the QOL-RTI\textsuperscript{97} and the EORTC-QLC C30/H&N 35,\textsuperscript{96} as well as several general QOL and psychological scales. Relationships have generally been in the expected directions.

One area of weakness for the FACT-H&N is a relative lack of data on responsiveness. A single study which assessed 21 newly diagnosed patients with laryngeal cancer at 4 points in time, over 6 months from diagnosis,\textsuperscript{38} showed only a trend to improved QOL with time from treatment. However, it must be recognized that at present, very limited responsiveness data are available for any of the head and neck specific health questionnaires.

2.5 Health Utility Measurement

A. Requirements for a Utility Measure in this Study

For this study, a global indicator of the value of an individual’s health state was needed. Measurements were to be carried out on actual patients who were experiencing their own health
2. Background

states after laryngeal cancer treatment. For this reason, it was not necessary to construct scenarios describing symptoms or functional ability; each participant simply rated his or her own situation. We chose to use a method intended to be understandable to the population of interest, which consists mainly of older males often with relatively limited levels of education. It was also important that the method be conceptually similar to the chronic nature of patients' conditions. For practical reasons, the method needed to be relatively quick to complete in a busy clinic setting.

Keeping in mind these ideal characteristics, it was also important to choose a method with acceptable reliability and validity data. In particular, a method that had been validated in cancer patients was preferred. The method needed to have been validated as a discriminative instrument, but responsiveness data was also desirable. As with the health questionnaire, we preferred to choose a method of utility measurement which has been used and recommended for patients with laryngeal cancer, so that our MID results might prove more useful to future researchers.

B. Choice of Utility Instrument for this Study

Three major methods of utility assessment have been described: the standard gamble (SG), time trade-off (TTO) and the rating scale (RS). Other potential methods, which include the equivalence method and ratio scaling have been less generally accepted and were not considered for our study. Probability trade-offs have been used frequently in assessing treatment preferences in cancer, but do not intuitively lend themselves to patient assessment of their personal utilities long after the treatment choice has been made.

The standard gamble (SG) is a method in which participants are asked to choose between an intermediate sure outcome, such as the present health state, or a gamble involving the best and
2. Background

The probabilities of the gamble are varied until the individual is indifferent about whether to gamble or to take the “sure thing.” It is derived directly from expected utility theory and is therefore valid by definition. Although it is generally regarded as the “gold standard” method of utility measurement, it requires an understanding of probabilities and can be difficult for patients to understand.

Unlike the SG, the time trade-off (TTO) was developed specifically as a simple measure for use in determining health utilities. It can be shown by utility theory to be equivalent to the SG under certain conditions. These conditions require that time is taken into account when using the SG, and risk is taken into account when using the TTO. Then, if the patient values each consecutive year of life equally, utility values measured by the TTO technique should match those measured by the SG. The mean is the most appropriate aggregation measure of central tendency for the TTO.

The rating scale (RS) differs from the SG and TTO in that it cannot be derived from expected utility theory. It is instead a direct rating by the patient of the health state of interest, and is free of risk or time considerations. In addition, it does not require a choice, and so may also differ in that regret would not play a part in ratings derived by this method. The reliability of these methods has been assessed using both internal consistency and test-retest approaches. Torrance reviewed the subject and found coefficients of internal reliability ranging from 0.77-0.92 for the SG, 0.77-0.88 for the TTO, and 0.70-0.94 for the RS. For test-retest reliability, the values were 0.80, 0.63-0.87, and 0.77 for the respective measures. For group comparisons, reliability coefficients greater than 0.50 are acceptable, whereas for comparisons between individuals, coefficients above 0.80 are desirable.

In practice, the TTO has been validated against the SG. Although correlation between SG and TTO ranged from 0.65 to 0.84, the TTO method gave utility values which were
systematically slightly lower. This fact has been attributed to non-linearity of the marginal utility, which manifests as risk aversion with the SG, or discounting of future life years with the TTO.\textsuperscript{107, 111} Essentially, this means that most individuals will value years in the near future more highly than those in the distant future, even if their health state remains the same. This effect has been demonstrated in cancer patients.\textsuperscript{112} Since adjustment of the TTO by a certainty equivalence factor designed to correct for non-linear marginal utilities failed to completely abolish the differences between TTO and SG values, a gambling effect (simple aversion to the act of taking a risk) may also play a role.\textsuperscript{109, 113} However, the TTO is generally easier to complete, and has been recommended for use in cancer patients, particularly in clinical situations where immediate death is not a likely consequence.\textsuperscript{114}

In attempting to establish criterion validity for the RS method by comparison with the SG and TTO, Torrance\textsuperscript{104} found that a power transformation was required to make either SG or TTO utilities comparable to RS scores. The relationship $RS = 1 - (1 - \text{utility})^x$ has been confirmed, with the $x$ value ranging from 0.47 to 0.64.\textsuperscript{107, 108} Correlation between transformed group TTO means and RS scores was 0.65 in one study.\textsuperscript{110} For individual patients' utility values it was not possible to define a transformation to convert TTO values to RS scores.\textsuperscript{104, 107}

We felt that the uncertainty that persists in the transformation function for the RS method argued strongly against its use to measure utilities unless practical considerations made other methods impossible. In considering the alternative methods of SG or TTO, we were influenced by the fact that the TTO is simpler for patients to understand, has been used more extensively in cancer patients, and avoids gambling, which would not come naturally to patients in stable situations who are not immediately facing the prospect of death. We selected the TTO as our established utility measurement technique.

The TTO technique has been used in three\textsuperscript{30, 69, 70} of five published studies\textsuperscript{30, 67, 68, 69, 70} assessing
utilities in larynx cancer scenarios.

C. An Experimental Utility Technique

The application of a numerical value to an underlying utility with the TTO instrument assumes linearity of the value placed on time. However, as discussed above, this assumption is unlikely to hold true because a patient’s values may change over time or with age, and remote life years are often valued less highly than are those in the near future. It may also be difficult for a patient to judge the value that he or she will place on remote periods of time. Decisions about one’s entire life are infrequently made, which may increase the complexity of the task. Moreover, the use of a patient’s remaining life span as the currency of exchange for the TTO may be a distasteful reminder of the inevitability of death, and may be particularly sensitive for cancer patients. Finally, the TTO implies an irreversible decision. Just as some risk-averse individuals may indicate artificially high utilities using the SG, individuals may over-rate their true utilities with the TTO because they are averse to making a permanent decision.

All of the above considerations led Buckingham to develop an alternative daily formulation of the TTO, in which patients are asked to give up a portion of their active, waking time each and every day for the rest of their lives in order to improve their health. Additional dreamless sleep, from which the respondent will awaken no more refreshed than usual, is used as a metaphor for the time thus given up. We have termed this method the Daily Active Time Exchange (DATE). In a comparison with the TTO in a large sample of Scottish adults (n = 1956), the DATE showed a higher response rate, a greater willingness on the behalf of respondents to trade at least some time, and better construct validity. The mean utilities were 0.84 for the TTO and 0.82 for the DATE. The correlation coefficient between the two methods was moderate but significant at 0.37. Construct validity was demonstrated by the fact that the
DATE correlated more highly (0.22 versus 0.12) than the TTO with a “health factor” derived from the SF-36.

D. Anchors and Time Interval

Utilities are measured on a scale of 0 to 1, in which 1 represents the most desirable and 0 the least desirable health state. The usual anchors are “perfect health” and death. However, several authors suggest the existence of states worse than death,116, 117, 118 for which the measurement of utilities is problematic.119 Fortunately, there is no evidence to suggest that such utility values would be expected in larynx cancer patients who are free of disease following treatment. We therefore accepted death as an appropriate lower anchor of our scale.

Consideration was, however, given to alternatives to the “perfect health” anchor. We were concerned that the concept might not seem realistic or imaginable to patients who had been treated for cancer. We entertained the idea of asking patients to trade time to achieve their “previous health,” defined as their state of well-being prior to noticing any symptom of laryngeal cancer. A study in angina used a similar approach with the SG.120 Since it is known that framing effects121, 122 and slight changes in anchoring112, 123 can significantly alter utilities, we rejected a TTO thus anchored as being of unknown validity and therefore inappropriate for use as a standard measure. We used the WHO definition of “perfect health” as the upper anchor for our assessments.4

The time period of consideration for the TTO is the respondent’s remaining life expectancy. When working with surrogate respondents, this is easily determined from life tables. However, for patients who may have multiple health problems and who have been treated for cancer, such a value is less easy to obtain. Moreover, we were concerned that it might be uncomfortable for
patients to be confronted with limitations on their remaining life. We have therefore followed the suggestion of Stiggelbout\textsuperscript{112} by asking patients to define their own expected remaining life span, and using this time period as the currency of exchange.

2.6 Estimation of the Minimal Important Difference

The minimal important difference was defined as the smallest change in value on a measurement instrument, which, from the point of view of the patient, represented an important rather than trivial change. In practice, it has been estimated for groups by the use of the minimal detectable difference, that is, the smallest difference which is detectable by the average patient.\textsuperscript{124}

Minimal important difference represents a heuristic\textsuperscript{125} or "anchor based"\textsuperscript{126} method of assessing instrument responsiveness to change. The alternative, a statistical or distribution based approach, is usually carried out on a group of patients in whom change is believed to have occurred. It involves calculating the responsiveness of an instrument with measures such as relative efficiency, measurement sensitivity, standardized response mean, effect size or the receiver operator curve.\textsuperscript{125} These measures help to determine if the instrument captures any change, but they do not measure or correlate to clinical or meaningful changes. Use of the MID has been proposed in order to invoke the patient's perspective.\textsuperscript{127} The different procedures for estimating MID which have been developed are discussed below.

A. Change over Time (Intra-Patient Method)

The intuitive approach for most physicians is to ask the patient whether a change has occurred.\textsuperscript{128} This is consistent with clinical practice. In groups, a transition score can be used to separate patients who indicate change from those who do not. The mean change in the subgroup indicating the smallest amount of change on the transition scale provides an estimate of the MID.
2. Background

This approach has been taken in COPD, asthma and cancer. It has even been attempted for N of 1 randomized trials. The approach has the advantage of direct applicability, since the measurement is taken in roughly the same way that it will be applied. It may also be the most appropriate approach for measuring change in highly personal or sensitive areas, such as sexuality.

However, the disadvantages of the intra-patient method include the need for prospective longitudinal assessment, and the potential for poor validity due to recall bias. Depending on the clinical situation, it may also require large sample sizes to include sufficient numbers of patients who have changed. It may present difficulties in estimating MID for patients who have worsened, since fewer people may report negative change. It is also most suitable for determining MID for instruments which have been shown to be responsive to change over time; in general, most QOL and utility instruments have been more thoroughly assessed for their ability to discriminate between individuals at a single point in time than they have been for their ability to respond to change.

The use of transitional indices, or change scores, has been criticized by Norman. He points out that reliability and validity studies are lacking for transition scores, and refers to the psychological difficulty of judging change due to recall bias and anchoring on the present state. He also notes that measurement errors on the measure of interest are unlikely to be independent from errors on the transition scale. He also argues that it may not be valid to combine the estimates of MID for patients who report having improved and worsened. In addition, he demonstrates mathematically that a coefficient of responsiveness obtained through such a study will always be greater than zero, due to the distribution of change scores in the sample, even if no average change has occurred in the cohort. Thus such a coefficient may overestimate the responsiveness of the measure, as compared with a statistical or distribution approach.
2. Background

B. Differences Between Individuals (Inter-Patient Method)

Another approach is to estimate MID by measuring the minimal difference that patients can detect when comparing themselves to others. The procedure involves having patients meet in groups and subjectively rate themselves on the construct of interest, as compared to other patients with the same diagnosis. This design was pioneered by Redelmeier using a functional status instrument in arthritis patients,\textsuperscript{138} and has since been used for a global assessment and pain score in arthritis,\textsuperscript{139} and, in COPD, for spirometry, dyspnea,\textsuperscript{140} the 6-minute walk,\textsuperscript{141} and a disease specific QOL instrument.\textsuperscript{128} This approach offers several advantages. Patients do not need to be followed prospectively; nor is it necessary for treatment-related change to have occurred. The MID for both positive and negative change can be measured, and the validity of inter-personal change can be validated against self-rated measures. It is an explicit measure in which patient bias can be detected. Smaller sample sizes may be possible since a single individual generates multiple comparisons.

However, some of Norman's concerns may also apply to the inter-patient approach. Reliability and validity have not been directly tested for global subjective comparison ratings, and again, measurement errors for the self-rated QOL instrument and the subjective comparison may not be independent. In this method, however, positive and negative MID scores can be calculated separately and are pooled only if they are statistically similar. The mathematical argument regarding overestimation of responsiveness using change scores probably does not apply, since no actual change is allowed to take place in the inter-patient method. Furthermore, the approach avoids recall bias and offers practical advantages with lower sample size and applicability to discriminative instruments.

Internal reliability of the inter-patient method has been shown by the comparison of MID estimates generated by an entire study sample, and by various subgroups who might be expected
2. Background

to differ.\textsuperscript{138, 141} For example, in COPD patients the MID for the 6-minute walk was estimated as 54 m in a group of 112 patients in a respiratory rehabilitation programme.\textsuperscript{141} No significant difference from this value was seen in subgroups generated according to gender, age, level of FEV-1, and baseline functional status.

Establishing validity for methods of MID estimation is problematic. Criterion validity cannot be established in the absence of a gold standard, and no such criterion exists for MID. Face validity refers to the degree to which a method makes sense to its actual and potential users. Traditionally, physicians have used informal transition indices whenever we asked patients, "are you feeling better today?" and so the face validity of the intra-patient method is high. However, in clinical trials we routinely compare patients with each other, supporting the face validity of the inter-patient method. Construct validity requires the establishment of hypotheses about the methods, and confirmation of these hypotheses in studies. For example, we would expect the mean transition score or mean subjective comparison scores to correlate with mean arithmetic differences on the QOL scores. In his 6-minute walk study, Redelmeier found a significantly lower correlation coefficient of 0.20 for an intra-patient transition index versus 0.59 (p < 0.001) for an inter-patient subjective comparison rating.\textsuperscript{141} Similarly, given a constant total score, the expected minimal detectable difference might be lower for a multi-item instrument than for a single item instrument due to better expected responsiveness.\textsuperscript{142} Since MID has not yet been measured for a multi-item scale of QOL with a summary measure, this relationship has not yet been fully assessed. MID for the 12-item activities subscale of the Asthma Quality of Life Questionnaire was 8.0\% or 0.56 per question.\textsuperscript{129} However, for the 8-item Stanford Health Assessment Questionnaire (HAQ) functional status instrument in arthritis patients, MID was 6.3\% or 0.51 per item in one study,\textsuperscript{138} while another found a 7.2\% positive and 16.2\% negative MID.\textsuperscript{139}

Concurrent validity could also be demonstrated by comparing MID results using two different
methods of assessment. This has been done in a study in which inter-patient estimates of MID for QOL in COPD patients were compared with historical estimates from the intra-patient method. The values obtained were remarkably similar for the entire instrument, and for 3 of 4 scales assessed. However, because of the retrospective nature of the study, tests of significance were not used. Certainly, additional evidence of validity would be provided for both methods if the same group of patients could be induced to undertake the inter-patient assessment prior to treatment, then be assessed prospectively with the intra-patient method. However, as the relative patient burden for each of the methods is rather high, such a study design would be extremely difficult in practice.

C. Applications to Cancer Patients

We are aware of only one published study of MID in cancer patients. Using the intra-patient method, Osoba administered the EORTC QLQ-C30 at baseline and at 2 follow-up intervals to 246 breast cancer patients and 80 small cell lung cancer patients undergoing chemotherapy in two randomized trials. Patients also completed a 7-option transition scale, the Subjective Significance Questionnaire (SSQ) at the follow-up assessments. For the group, no overall change in physical, social or global QOL scores occurred, but emotional function scores did improve from baseline. Patients rating themselves "a little better" had mean score changes of 6.9-18.1, except for the physical function score in breast cancer patients, in which no change was observed. Patients rating themselves "a little worse" had mean score changes of -9.6 to -16.7, except for the emotional function score in breast cancer patients, in which the mean change was 8.0. Correlations between the SSQ and corresponding QLQ-C30 subscales ranged from 0.23 to 0.45. Linear regression looking at different patient groups and subscales suggested a 4.8% to 12.2% percentage estimate of MID for a one-category difference on the SSQ.
D. Choice of Method of Estimating MID

We agree with other authors that the only sensible way of estimating the MID is by asking patients. Therefore, "anchor type" measures must be used. The inter-patient method offers several theoretical and practical advantages over the intra-patient method, and we judged that QOL for disease-free laryngeal cancer patients would not be of a sufficiently private or sensitive nature to interfere with its use. We also recognized the feasibility of this approach given a clinical setting in which a large number of laryngeal cancer patients are attending for follow-up. Given that the inter-patient method has not yet been applied to cancer patients, we saw the opportunity to produce novel results by choosing this approach for our study.
3. METHODS

3.1 Population

A. Potential Participants

This project was carried out as a single institution study at Princess Margaret Hospital. Patients were accrued over a 6 month period between May and November of 1998. Patients attending for follow-up of laryngeal cancer were identified, and a standardized letter [Appendix B] introducing the study was sent out 4-6 weeks prior to each appointment date. Patients were given a contact number to call if they preferred not to discuss the trial at their visit. A single investigator (GJR) attended all clinics and interviewed potential participants.

B. Inclusion Criteria

Patients were eligible for the study if they had squamous cell carcinoma of the larynx, stage T1-T4N0-3M0. They were required to have completed curative treatment for the disease at least 6 months prior to study enrollment, and to be free of active disease when assessed. Age greater than 18 was required, but in practice, Princess Margaret Hospital is not a pediatric centre and no such patients attended the clinic during the study period. Individuals were approached regardless of the mode of cancer treatment and regardless of whether or not they had experienced previous relapse.

Occasionally, it was not possible for the investigator to speak to all the potential patients in a busy clinic. At such times, patients were captured at their next follow-up visit. However, 24 potential patients did not attend for a second visit during the study period and were not approached for the study.
C. Exclusion Criteria

At the initial interview, patients were excluded if they were deemed incapable of participating in the group meeting. Thus, patients were required to speak and read English sufficiently to complete the instruments, and to converse about their disease. Blind, deaf or mute patients were similarly deemed ineligible. One patient who wished to participate was fluent in English, but illiterate. He was able to participate because his wife attended the group meeting with him to read response options aloud and record his responses. Patients who developed a recurrence between their initial visit and the group meeting were withdrawn from the trial, since it was anticipated that the recurrence might cause a change in QOL. For patients who preferred not to enter the trial, a reason was elicited and categorized.

3.2 Baseline Variables

A. FACT-H&N Health Questionnaire

At the initial clinic visit and following instruction, participants were given the opportunity to complete the FACT-H&N [Appendix D] privately without interruption. Where sought, clarification on the questions was provided by the investigator. This consisted of reading the question to the patient, ascertaining his area of uncertainty, and attempting to assist without biasing the response. For example, many patients were estranged from or no longer had living family members, and were uncertain of how to answer the questions regarding family on the social subscale. In such instances, they were instructed to think of the people they considered to be their family, or, if no one in their lives fit that description, to leave the question blank.
3. Methods

B. Utility Measures

The utility instruments were administered after the FACT-H&N, so that each patient was given the opportunity to think about his or her personal situation prior to being asked to rate it on the TTO [Appendix E] and DATE [Appendix F]. As order effects have been previously observed,\textsuperscript{115} the order of administration of the two utility instruments was pseudo-randomized by creating equal numbers of non-sealed interview packages with one or the other instrument placed first. Both utility instruments were administered in a structured personal interview with the investigator (GJR). Patients were permitted to have a friend or family member present during the interview, but were asked to reply according to their own thoughts and feelings.

The administration technique consisted of reading the text to the patient, with repetition of any areas that seemed unclear. “Perfect health” was described using the WHO definition.\textsuperscript{4} As discussed in section 4.2.A, this concept presented difficulties to many patients. In the TTO, patients were asked to estimate how much longer they expected to live. This often elicited considerable reflection, and patients were encouraged to express their thoughts. With the exception of one individual, all participants were able to provide an estimate. No judgements were made regarding plausibility, but no patient expressed a nonsensical estimate. For the DATE, patients were asked how long they slept in an average day, and no one had difficulty providing this time period.

For both the TTO and the DATE, the “ping-pong” method\textsuperscript{103} was used to avoid anchoring effects. As a TTO example, a patient expecting to live 10 years would first be offered the options of 10 years in his present health, versus 1 year in perfect health. Assuming he rejected the one year option, he would be offered 10 years in his present health versus 9 years in perfect health. The years in perfect health would continue to be varied in the pattern 2, 8, 3, 7, etc. When patients reached a point of indifference, the number of years or hours in perfect health was divided by the
number of years or hours in present health to obtain the utility score. For example, if a patient indicated that being awake and in perfect health for 12 hours a day was equivalent to being awake and in his present health for 16 hours a day, that patient would score 12/16 or 0.75. If a patient was unable to clearly determine a point of equivalence, the midpoint between the years or hours that he would and would not trade was used to calculate utilities. For example, if the above patient could not identify the point of equivalence but was willing to be awake for only 12 hours and was not willing to trade if offered 11 hours, his score would be 11.5/16 or 0.72. Patients who indicated that they would not trade any time to achieve perfect health were given a score of 1.0. For patients scoring 1.0, a distinction was made between those who felt their health was already perfect, those who did not want perfect health, and those who did not have and would like to have perfect health, but were not willing to give up any time to achieve it.

C. Demographic, Tumour and Treatment Factors

Chart review was undertaken for all eligible patients, regardless of whether or not they participated in the trial. Baseline data included patient age, disease stage, grade and subsite, treatment type and completion date, number of relapses, time since initial diagnosis, time since completion of initial treatment, and time since completion of last treatment. For simplicity, no attempt was made to identify or quantify comorbid illness.

D. Performance Status

Karnofsky performance status [Appendix G] at the time of clinic visit was recorded whenever possible for both participants and non-participants. We also wondered if the Karnofsky score at the time of initial diagnosis might predict QOL after treatment. Karnofsky score had not been collected prospectively for most patients, and so an attempt was made to impute Karnofsky score at diagnosis by asking the patient to remember and describe their situation when they first found
out they had cancer. In some instances it was not possible to impute a missing Karnofsky score: for example, in a non-English speaking patient when a translator was not available.

3.3 Group Meetings

Following the initial assessment, each patient was offered his choice of several dates for a group meeting. Organization and management of these meetings required considerable time and effort. Group assignment was carried out according to patient convenience, and actual group composition was further influenced by the willingness of some patients to attend. Non-attenders were contacted by telephone, and if they expressed a continued interest in the trial, they were rescheduled. Several non-attenders subsequently withdrew from the trial without ever attending a group meeting. When this occurred their reasons for withdrawal were recorded.

Meetings were held on Saturday or Sunday afternoons, between the hours of 1:00 and 5:00 p.m. They took place at the Princess Margaret Hospital Lodge, an accommodation facility for patients undergoing cancer treatment which provided complimentary parking and a comfortable, relaxing environment. For various group sizes, schedules were constructed a priori to permit interaction of every group member with every other member. On the day of the meeting patients were assigned nametag numbers corresponding to their order of arrival, and the appropriate schedule was chosen for the actual group size. Thus, we made no conscious effort to influence the make-up of patient groups or the order of patient interactions.

A. FACT-H&N Health Questionnaire

At the group meeting, patients were asked to complete the FACT-H&N for a second time. FACT-H&N was repeated for three reasons. First, we wished to identify patients for whom actual changes had taken place between the individual and group meetings. Secondly, we wished
to re-prompt each patient to think about his own QOL prior to comparing himself to others. Finally, test-retest reliability has not yet been published for the FACT-H&N, and administering the instrument twice allowed us to calculate the correlation between the two administrations. The utility instruments were not repeated at the group meeting because the atmosphere may have been relatively distracting with many others present in the room. Since patients had less time and privacy when answering FACT-H&N #2 at the group meeting, we used FACT-H&N #1 in all analyses unless otherwise noted.

B. Utility Measures

The utility assessments were not repeated due to time constraints and the difficulty of carrying out cognitively complicated assessments in a group. On two occasions, patients who wished to participate did not have time to complete the utility assessments when seen initially in the clinic. For these patients, time was set aside to complete the utilities prior to the group meeting and before the arrival of the other group members. One such patient withdrew from the trial prior to attending the second meeting, and therefore no utility values were available for him. Otherwise, all utilities were elicited in private during the earlier scheduled clinic visit.

C. Inter-patient Comparisons

Each patient was given a pen, copies of the Subjective Comparison Rating Form [Appendix H], and a nametag with his number. In a brief introductory lecture, health-related QOL was described. Patients were then instructed to speak one-on-one with their partners, and to discuss any issues they felt were important to their health or QOL. They were explicitly asked to avoid discussing extraneous topics such as the weather, the news, etc. They were asked to share descriptions of their current well-being and any problems they might be having as a result of their disease or treatment.
Each patient was serially paired with every other patient in the group. Each interaction lasted about 10 minutes. Following each pairing, patients separated and confidentially rated themselves compared to the people with whom they had just met. They were instructed that there were no correct answers, and that we were simply interested in their opinions. We also explained that it was not necessary for ratings to be reciprocal; that is, if A said he was a little bit better than B, then it did not matter if B also rated himself as a little bit better than A. They were strongly encouraged not to share their ratings. Rating forms were collected after each round, following which patients were paired with different partners and the process was repeated.

After each pairing, each patient rated himself compared to his partner on the Subjective Comparison Rating Form, which included 6 separate scales, designed to match the subscales of FACT-H&N. The scales rated comparisons of physical, emotional, social, functional, head and neck specific, and overall well-being. Each was a Likert scale with values from 1 (much better) to 7 (much worse). The value 4 represented “about the same”, with 3 and 5 representing “a little bit” better and worse, while 2 and 6 were defined as “somewhat” better and worse. This scale is standard for both intra-patient and inter-patient comparisons.127, 128, 129, 130, 138, 139, 140, 141

We were initially concerned that open discussion of QOL issues might be upsetting for the patients. We arranged for a social worker from the head and neck cancer team at Princess Margaret Hospital (Mr. Stanley Chan) to be present during the first 4 meetings. However, since no problems occurred he did not attend the final 6 meetings. Occasionally, patients were not able to stay for the entire meeting. Due to this fact, data was not available for a few of the planned comparisons.
3.4 Unanticipated Changes

As described above, we attempted to identify patients for whom a change in FACT-H&N score occurred between the two meetings. We did so by directly comparing the two sets of scores. Two methods were used to assess change: a judgmental and a statistical approach.

A. Judgemental Approach

A scatter plot was constructed comparing scores on FACT-H&N #1 and FACT-H&N #2 for individual patients. Visual inspection was used to identify patients as potential outliers with a more distinct change in score than others in the study. The plot is shown for our data in Figure 1.

B. Statistical Approach

Previous data on the FACT-H&N indicates the instrument standard deviation is about 16 for FACT-G and 7 for the H&N subscale. We took the square root of the squares of these values to calculate a standard deviation for FACT-H&N of about 17. We defined an outlier as any patient for whom the two FACT-H&N scores differed by greater than 2 standard deviations:

\[ | \text{FACT-H&N} \ #1 - \text{FACT-H&N} \ #2 | > (2 \times 17.45). \]

C. Management of Outliers

Both visual inspection and the statistical approach yielded the same result, with one patient identified as an outlier. Serendipitously, that individual had completed her utility assessments on the day of the group meeting. We therefore chose to include her results but use her FACT-H&N #2 for comparisons with utilities and subjective comparison ratings. Hence, no data was excluded due to unanticipated changes in health status.
3. Methods

3.5 Statistical Analysis

Statistical analysis was carried out using STATVIEW software on a Macintosh Performa 580 computer. All data was double-entered and inconsistencies were reconciled by checking the original hard copies.

A. Sample Size

Sample size, based on previous studies using similar methodology, was planned to include 5 groups of 20 patients each, for a total of 100 patients. Each pair generated two contrasts, that is, A's rating compared to B, and B's rating compared to A. The unit of analysis for correlation and MID quantifications was the contrast rather than the patient. Each group would have provided 380 contrasts for a total of 1900 contrasts. Assuming the above, this would have provided 80% power to detect a correlation coefficient (between arithmetic differences and subjective comparison ratings) of 3% or higher with a p value of 0.05. Smaller coefficients would explain about 0.1% of the total variance and were judged to be clinically insignificant.

In reality, not all scheduled patients attended their group meetings, and we found the larger groups to be tiring for participants. Thus, in the study, 98 participants met in 10 groups ranging from 7 to 16 participants. Each group provided between 56 and 182 contrasts, for a total of 938 possible contrasts. Due to incomplete data, 823, 847, 857, and 861 subjective contrasts could be used respectively in comparison with differences of TTO, DATE, FACT-H&N and Karnofsky score. Post-hoc power calculation suggests that this should provide adequate power to detect correlations (between arithmetic differences and subjective comparison ratings) of 10% or higher, which explain about 1% of total variance.\textsuperscript{144}
B. Analysis of Minimal Important Difference

For each paired comparison, the arithmetic difference between the two individuals' scores was calculated for FACT-H&N, TTO, DATE, and current Karnofsky score. These values were coded as DIFF-FACT-H&N, DIFF-TTO, etc. For example, if a patient with a FACT-H&N score of 100 interacted with a patient with a FACT-H&N score of 70, the DIFF-FACT-H&N was 30. A scattergram was used to visually inspect the degree of linearity in the relationship between DIFF-score and COMP-overall, the patients' subjective comparison of overall well-being.

As the individual data points were expected to be somewhat scattered, it was necessary to calculate mean arithmetic differences for each level of subjective comparison rating. For example, for patients who rated themselves as "about the same" (COMP-overall = 4, on a seven point scale), we calculated the group mean for each of the DIFF variables. We undertook the same procedures for all levels of COMP-overall, and for all of the DIFF variables. We were then able to calculate an estimate of MID. For example, for the TTO, the mean DIFF-TTO for COMP-overall = 4 ("about the same") was subtracted from the mean DIFF-TTO for COMP-overall = 3 ("a little bit better") to obtain an estimate of MID for the TTO. A second estimate of the MID was calculated by a difference of means for patients rating themselves as "a little bit worse" rather than "about the same." Student's t-test was used to compare these two "positive" and "negative" estimates of MID, and if no significant difference was found, the average value was reported.

Using the same process described above, we estimated the MID for the six subscales of FACT-H&N by comparing arithmetic difference and subjective comparison scores matched for each dimension. This insured that we were comparing similar issues for each analysis. For example, to estimate the "positive" MID for the physical subscale (FACT-P) of FACT-H&N, we obtained the difference between the mean DIFF-FACT-P for patients rating their physical well-being as "a
little bit better" versus "about the same" on COMP-P (subjective comparison of physical well-being).

C. Test of Robustness

To check the reliability of our data, we dichotomized patients according to age, gender, relapse status, and laryngeal subsite and repeated the above process to estimate MID for each subgroup. We plotted means and 95% confidence intervals to assess whether estimates of MID differed according to any of these potential predictive variables.

To assess the degree to which our estimates of MID derived from the whole group of participants could be applied to individual patients, we plotted the cumulative probability of rating oneself as at least "a little bit better" for several levels of arithmetic difference in FACT-H&N, TTO, DATE, and Karnofsky scores. A previous study using the inter-patient method of estimating MID\textsuperscript{140} has suggested that individual variation is sometimes too great for group estimates of MID to predict how individual patients would judge their own QOL.

To support the validity of the subjective comparison ratings used in this trial, we also measured the degree of correlation between the arithmetic differences (DIFF) and subjective comparison ratings (COMP). Spearman correlation coefficients were reported with P-values.

D. Relationship Between Measures of QOL

We used Spearman correlation coefficients to analyze the relationship between FACT-H&N, TTO, DATE, and Karnofsky. The significance of all observed correlations was determined by reporting P-values.
E. Generalizability

To determine the extent to which our sample was representative of all patients being followed for laryngeal cancer at the Princess Margaret Hospital, we compared the patient, tumour and treatment variables of our participants to those of eligible patients seen during the trial period who were excluded or who chose not to participate. Student’s t-test was used for all comparisons. We recognize that for multiple comparisons, the conventional P-value of 0.05 may be insufficiently rigorous to exclude a difference due to chance. However, as we wished to avoid missing a true difference between our sample and the population in order to be conservative regarding the generalizability of our results, we did not adopt a lower P-value.

3.6 Ethics

A. Informed Consent

The study was explained by the investigator (GJR), who in most instances was not a member of the patient’s health care team. Every effort was made to insure that patients realized the voluntary nature of trial participation. No monetary or other incentives to participate were offered. Patients who chose to participate were required to read and sign a consent form [Appendix C]. If any question existed regarding a patient’s competence to consent, that patient was excluded from the trial.

B. Impact on Patient Care

The study did not impact in any way on patient management. No material risks were incurred by participants. We made every effort to communicate to patients that a preference not to participate would in no way influence their relationship with their health care team.
3. Methods

C. Data Management and Reporting

Raw data sheets were maintained in locked offices. All data entry and analysis was carried out in locked facilities using password protected computers. This and any following scientific reports will discuss aggregate statistics only, with no identification of individuals.

D. Ethics Committee Approval

Approval to conduct this trial was sought and granted by both the Princess Margaret Hospital Clinical Trials Committee [Appendix I] and the Oncology Human Review Committee at the University of Toronto [Appendix J].
4. RESULTS

4.1 Recruitment

A. Degree of Participation

Over the 7 month recruitment period, 339 patients met the study inclusion criteria. At clinic visits, 315 were approached for the trial. In total, 70 of these 315 patients were excluded according to pre-determined criteria, leaving 245 eligible candidates. The trial acceptance rate was 49%, with 121 individuals accepting and 124 declining participation. Of the 121 patients enrolled, 23 subsequently withdrew prior to attending a group meeting. Of these 23, one patient also did not complete the utility component of the baseline assessment. Therefore, complete baseline assessments were available on 120 patients, and group comparisons on 98. In this report, baseline results have been presented for all 120 patients but MID analysis includes only the 98 patients completing the entire study requirements.

B. Reasons for Patient Exclusions

Overall, 70 patients who were initially interviewed were excluded from participation. Lack of English language skills accounted for 60 exclusions. An additional three patients were illiterate, three were blind, one was deaf, one was mute, two were not competent to consent due to Alzheimer’s disease and one was in a state of alcohol intoxication at the clinic visit.

C. Reasons for Declining Participation

Participation in this study involved an additional visit to Toronto on a weekend for a group meeting. Given the large referral base of the hospital, this entailed travel over long distances for
many of our patients. The study also required a significant time commitment and willingness to share personal information. Given these considerations, the study acceptance rate of 49% was close to our prediction of 50%.

Of the 124 patients declining the study, 50 stated that their main reason was the distance, and 37 indicated the time commitment. Discomfort talking with others about their illness was cited by 23, while poor health due to comorbid illness prevented 11 patients from taking part. Finally, three individuals did not state a reason.

D. Reasons for Withdrawals

After enrollment, 23 patients withdrew prior to the group meeting. One patient died suddenly of cardiac causes, and one developed a recurrence of laryngeal cancer. Eleven patients decided they did not have time to participate, and three decided the distance was too great. Two patients reconsidered and felt uncomfortable talking with others. Three patients were unable to attend due to worsening of comorbid illness, and one could not attend due to an illness in his family. One individual failed to attend his scheduled group and was subsequently lost to follow-up.

E. Representativeness of the Sample

In Table 1, baseline patient, disease and treatment factors for 98 participants and 241 non-participants (24 not approached, 70 exclusions, 124 decliners and 23 withdrawals) are compared. The 98 study participants were somewhat younger (mean age 65 versus 68), and had a slightly higher Karnofsky performance status (mean 88 versus 84) than the 219 non-participants. No differences were observed on other potential predictive factors.
4. Results

4.2 Baseline Results for QOL and Utilities

A. Qualitative Results

All patients were able to complete the FACT-H&N. No questionnaires were excluded from analysis due to missing responses.

Utility assessments were made in 114 patients. One patient could not do the TTO because he was unable to estimate his remaining lifespan, saying, “When the time comes, it comes.” An additional patient found both the TTO and DATE tasks to be too abstract, was unable to complete the assessment, and was also among a group of 18 patients who stated that they did not want perfect health. The full list of patients’ reasons for declining perfect health appears in Table 4. Eleven patients felt that their present health was already perfect, and that they would therefore achieve nothing by trading any time.

When patients who could not do the task or who already had or did not want perfect health were excluded, TTO and DATE results, respectively, were available on 84 and 85 patients. The rate of willingness to trade at least some time was 46/84 (55%) on the TTO and 64/85 (75%) on the DATE. This represented a significant difference (p = 0.0027).

B. Quantitative Results

Mean baseline scores on FACT-H&N (123/144), TTO (0.91/1.0) and DATE (0.89/1.0) for all patients are presented in Table 2. In general, QOL scores were skewed toward the upper part of the total range for all three instruments. In Table 3, means are re-calculated for the TTO (0.89/1.0) and DATE (0.85/1.0), excluding those patients who stated that their health was already perfect, or who did not desire perfect health, since the utility instrument may not be valid in such
patients. Excluding these patients produced significant differences in both TTO and DATE (both $p < 0.0001$).

C. Correlation Between QOL Measures

A Spearman correlation matrix between the two administrations of the FACT-H&N (#1 and #2), TTO, DATE, and Karnofsky is shown in Table 5. All instruments were significantly correlated with P-values between 0.003 and 0.0001. None of the instruments correlated with patient age (data not shown). A reassuringly high correlation of 0.82 ($p = 0.0001$) was observed between the two administrations of FACT-H&N, suggesting good test-retest reliability and little interval QOL change. The health questionnaire was more strongly correlated with performance status ($r = 0.42$, $p = 0.0001$) than with utilities (TTO, $r = 0.30$, $p = 0.0017$; DATE, $r=0.34$, $p = 0.0004$). A moderately high correlation of 0.59 ($p = 0.0001$) was noted between the DATE and the TTO.

D. Clinical Predictors of QOL Scores

Univariate and multivariate regression showed gender to be the only significant predictor of FACT-H&N score ($r = 0.23$, $p = 0.01$), with males reporting better QOL (mean 125 versus 115). Univariate regression suggested number of relapses ($r = 0.23$, $p = 0.01$) and mode of speech ($r = 0.19$, $p = 0.05$) to be predictive for better QOL on the TTO, but in multivariate analysis only number of relapses remained significant ($r = 0.25$, $p = 0.04$), with mean TTO scores of 0.89 in patients with no history of relapse, versus 0.88 in patients with prior relapse. Laryngeal subsite was a predictor for DATE ($r = 0.19$, $p = 0.01$), with better QOL in patients with glottic rather than supraglottic tumours (mean 0.90 versus 0.85).
4. Results

E. Discrimination Between Known Groups

We wondered whether either health questionnaire or utility instruments were able to discriminate between treatment groups. We calculated mean values for patients who had and had not been treated surgically. Few patients were in the surgical group (n = 10). Although mean scores were lower in surgical patients, Student's t-test failed to show a statistically significant difference between groups for mean scores on FACT-H&N (120 vs. 123), TTO (0.83 vs. 0.91) or DATE (0.85 vs. 0.89). Borderline significance was seen on the TTO (p = 0.07).

4.3 Subjective Comparison Ratings

A. Analyzable Date

Ten group meetings took place, three with 8 participants, three with 9 participants and one with each of 16, 14, 10, and 7 participants respectively. The maximum possible number of contrasts was 938. After consideration of missing data, 861/938 (92%) of contrasts were analyzable for FACT-H&N, 823/938 (88%) for TTO, 847/938 (90%) for DATE, and 861/938 (92%) for Karnofsky.

B. Distribution of Differences and Comparison Ratings

The distribution of arithmetic differences in FACT-H&N was symmetric about zero, as would be expected given the reciprocal nature of these pairings (Figure 2). For example, if a patient with a FACT-H&N score of 100 met with a patient whose score was 70, the arithmetic difference in FACT-H&N was 30 for the first patient and -30 for the second. Similar distributions were seen for the differences in TTO and the differences in the DATE (data not shown).
In contrast, the distribution of subjective comparison ratings was highly skewed, as can be seen in Figure 3. The majority of individuals rated themselves as “about the same” as their partners. More individuals rated themselves as at least “a little bit better” than their partner than rated themselves as being worse. Indeed, not a single “much worse” rating was seen for the comparison of overall well-being.

C. Qualitative Observations

Scatter plots were constructed by plotting the arithmetic differences against subjective comparisons for the matching dimensions (Figures 4 and 5). The best-fit regression line showed the expected association, but individual values were widely distributed and not all data points were near the best-fit line. Similar scatter plots were seen for all comparisons of arithmetic difference and subjective comparison (data not shown).

Data was grouped as mean arithmetic DIFF values for each level of subjective COMP. Using this method, the relationship between arithmetic differences and subjective comparison ratings is shown in Figures 6 through 8. Figure 9 shows such a comparison relating difference on the head and neck subscale of FACT-H&N to the subjective comparison rating of head and neck symptoms (COMP-H). Similar comparisons were made for the other subscales of FACT-H&N, and for FACT-G. Correlations between arithmetic differences and subjective comparison ratings for the subscales of FACT-H&N are shown in Table 6.

4.4 Estimates of Minimal Important Difference

A. Correlation for Arithmetic Differences and Subjective Comparisons

Table 7 shows a Spearman correlation matrix for arithmetic differences versus subjective
comparisons (COMP-overall). Correlations are low to moderate at -0.19 to -0.22 and all are significant with P < 0.0001. Negative correlations were expected, since for the arithmetic differences a positive number indicated better QOL whereas for the subjective comparisons, a lower subjective rating score indicated better well-being.

Correlation of the FACT-H&N subscale differences with their matching subjective comparison ratings produced similar results. The highest correlation, as was expected, was for the most specific area, head and neck specific concerns (r = -0.33). The functional subscale showed a correlation of -0.22 as compared to -0.21 for the physical and -0.14 for the social subscale. All of the above correlations were significant with P < 0.0001. In contrast, no significant correlation was seen between differences on the FACT-H&N emotional subscale and subjective comparison ratings of emotional well-being (Table 6).

B. Asymmetry of MID

Both positive and negative estimates of MID were calculated. For example, the “positive” MID is the difference between the mean TTO among patients rating themselves as “a little bit better” than their partner minus the mean TTO among patients rating themselves as “about the same” as their partner. The “negative” MID is the difference in mean TTO scores of those rating themselves as “about the same” and “a little bit worse.”

In several previous studies,\textsuperscript{128, 138, 140, 141} no significant difference was noted between “positive” and “negative” MID, and the two values were averaged. In one arthritis trial,\textsuperscript{139} a larger “negative” than “positive” MID was found. Our results were in keeping with this final trial, and we have reported “positive” and “negative” MID estimates separately.
4. Results

C. “Positive” and “Negative” Estimates of MID

Table 8 shows estimates of MID derived from the subjective comparison of overall well-being. Table 9 shows MID estimates for the subscales of FACT-H&N, with each scale compared to its specific corresponding subjective comparison rating. For each estimate, the “negative” MID is larger than the “positive” one, suggesting that a small improvement in QOL or functional status may be more noticeable to laryngeal cancer patients than a similar degree of deterioration.

For the emotional subscale of FACT-H&N, no significant difference was seen between positive and negative MID estimates. Differences in score on this subscale also failed to correlate significantly with subjective comparison ratings of emotional well-being (Table 6). Therefore, it is questionable whether our method was successful for this subscale. It has been suggested previously\textsuperscript{132} that the inter-patient method may be less useful for highly personal or sensitive issues, and emotional well-being may be such an issue.

4.5 Tests of Robustness

A. Subgroup estimates of MID

Figures 10 and 11 show “positive” and “negative” estimates of MID for FACT-H&N drawn from subgroup analysis. Patients were divided by age as older (greater than or equal to the median age) versus younger than the median age. Subgroups were also assessed for gender (male versus female), laryngeal subsite (supraglottic versus glottic) and relapse status (prior relapse versus never relapsed). No significantly different estimates were observed, but 95% confidence intervals were wide. Similar results were seen for subgroup analyses of MID for TTO and DATE (data not shown).
4. Results

B. Relationship Between Item Number and MID

We had hypothesized that an instrument with a greater number of items, which would theoretically be expected to be more responsive than a single item global index, might have a smaller MID. It is difficult to compare MID across items with different total scores, but we attempted to do so by normalizing all MID estimates as a percentage of total instrument score. Data are shown in this manner in Table 10. It can be seen that positive MID estimates ranged from 3.8 to 5.5% for global measures (TTO, DATE, and Karnofsky) and from 4 to 4.3% for multi-item instruments (FACT-G, FACT-H&N). Similar overlap was observed for percentage estimates of negative MID for global and multi-item measures.

C. Applicability of MID Estimates to Individuals

We plotted the cumulative probability of rating oneself as at least "a little bit better" on the subjective comparison of overall well-being versus the difference in score on FACT-H&N, TTO and DATE. Figure 12 shows the resulting curve for TTO. This shows that at the estimated positive MID of 0.05, just over half of our participants rated themselves as at least a little bit better than their partners. However, if the difference was at least 0.15, about three quarters of the ratings were at least "a little bit better." Similar curves were found for FACT-H&N and DATE. Therefore, some patients are more sensitive to a difference in QOL or utility score than are others, limiting the applicability of average estimates of MID to individuals. Similar findings have been reported by others.
5. DISCUSSION

5.1 Estimation of Minimal Important Difference

A. Magnitude of the Estimates

We assessed the threshold at which a difference in measured QOL was associated with a small but noticeable difference in perceived well-being for laryngeal cancer patients. Using an inter-patient method and comparing arithmetic differences in QOL with subjective comparison ratings, we found thresholds ranging from 4 to 10% of the possible span of each instrument.

For the 36-item FACT-H&N, with its minimum score of 0, maximum score of 144 and 5 point Likert scale, the positive and negative MID estimates of 6 (4%) and 12 (9%) correspond to a mean change of about 0.2 to 0.3 per question or one point on every third to fifth question. Even with reversals from a totally asymptomatic to an extremely symptomatic state, difference would have to develop on a minimum of 2 items (improvement) or 4 items (deterioration) to be noticeable by the average patient.

For the TTO, the MID estimate of about 0.05 (5%) corresponds to giving up 1 year from 20 remaining life years. For DATE, and assuming 8 hours of usual sleep daily, the positive MID of 0.05 (5%) and negative MID of 0.14 (14%) represent about 48 minutes and 134 minutes of additional daily, dreamless sleep respectively.

Previous estimates of MID using the inter-patient method have been made for other health questionnaires. For the 20-item Chronic Respiratory Questionnaire (CRQ) in COPD patients, MID was about 8%. For the 8-item Stanford Health Assessment Questionnaire (HAQ) functional status instrument in arthritis patients, MID was about 6% in one study, while
another found about a 7% positive and 16% negative MID.\textsuperscript{139} Other inter-patient studies have determined MID for the 6-minute walk and FEV-1, measures which do not have an absolute maximum score. For instruments that have maximum scores, we have expressed MID as a percentage of the theoretical range rather than the usual observed range, which is approximately 1/7 of the theoretical range (Dr. Donald Redelmeier. Personal communication. July 23, 1999). For reference, MID for the 6-minute walk was 40 m and a typical normal value is about 700 m, giving a MID of about 6%.\textsuperscript{141} Similarly, for the FEV-1, the MID was 112 ml or 6% of a typically normal 2000 ml.\textsuperscript{140}

Two studies using the intra-patient method have found MID estimates of about 7%.\textsuperscript{127, 129} For cancer patients, a percentage change of about 5-10\% corresponded with the MID for different subscales of the EORTC-QLQ-C30.\textsuperscript{130}

Considering the differences in patients, disease and instruments in which these MID studies have been done, the similarity of percentage change estimates of MID is remarkable. Our estimates for the FACT-H\&N and two utility instruments are well within the previously reported percentage range for other QOL instruments.

Our study is the first to measure MID for the FACT-H\&N. However, for the FACT-G, a preliminary estimate was made by administering the instrument 2 months apart in a group of 104 patients on chemotherapy for advanced lung, breast or colorectal carcinoma. Instead of using a transition index, Karnofsky performance status was measured at both time points and patients were grouped as “improved”, “no change” or “declined” according to performance status. A mean change of 6.8/108 (6.3\%) was found for “improved” patients, and 5.4/108 (5\%) for “declined” patients.\textsuperscript{2} This is quite similar to our positive and negative MID estimates for FACT-G, 4.37/108 (4\%) and 8.28/108 (8\%).
5. Discussion

We are not aware of any publications that have attempted to measure MID for utility instruments. Expert opinion, however, suggests that, "changes of 0.1 are probably clinically important." Our estimates of 5 to 6% for the TTO and 5 to 14% for the DATE are slightly lower than this prediction.

B. Asymmetry of Positive and Negative MID

Our study consistently showed that a deterioration must be larger than the corresponding improvement for it to be detectable to the average patient. This held true for both site specific (FACT-H&N) and general QOL measures (FACT-G), both utility measures (TTO and DATE), and the Karnofsky performance status scale. We found the same result on four of the five subscales of FACT-H&N. In subgroup analyses, for all instruments except the TTO, this asymmetry was present regardless of age, gender, subsite and relapse status. For the TTO, statistically significant asymmetry was not observed in older patients, women, those with supraglottic tumours, or patients without a history of relapse; however, our study was not sufficiently powered to yield reliable subgroup analyses.

One previous study using the inter-patient method presented similar findings. Using the Stanford HAQ in arthritis patients, Wells found a positive MID of 7% and a negative MID of 16%. Similar asymmetry was found for ratings of pain and global well-being. All differences were statistically significant. The authors hypothesized that these findings likely represented a true characteristic of patients' attitudes towards health. Other studies have shown similar trends but not always statistical significance.

In most studies using the intra-patient method, including the single trial involving cancer patients, positive and negative changes were not reported separately. In asthma patients, positive and negative MID were reported for four domains and no consistent trend
was observed between their magnitudes. However, most studies using the intra-patient method are poorly designed to test for possible asymmetries of the MID.

It is possible that the trend to a higher MID for negative differences may be due to ego bias. Patients tended to rate themselves as “better” rather than “worse” compared to others, so that fewer estimates of negative MID were possible. However, the contrast we observed between negative and positive MID is unlikely to be due to the small sample size. Conceivably, the negative MID values were larger because those patients who did rate themselves as worse were anomalous individuals with a negative self-impression, which led them to underestimate their QOL on the self-rated measures. However, this conjecture is implausible given the consistency of our findings on different measures and in all patient subgroups.

We wonder whether a small improvement in QOL may indeed be more meaningful to cancer patients than a small deterioration. Such a difference in value would be consistent with ego bias. If true, this has important implications for study design and results. A treatment which has a small average negative impact on QOL might still be very acceptable to patients if it offers other advantages. On the other hand, a therapy with the sole advantage of improving QOL might be worthwhile even if it offered a relatively small average benefit.

C. Validity and Precision of MID Estimates

To assess the reliability and robustness of our data, we compared estimates of the MID obtained from our entire cohort to those obtained for subgroups divided according to age, gender, relapse status and laryngeal subsite. We chose to construct subgroups according to these characteristics because previous studies and our regression analysis suggested that they might influence QOL. The confidence intervals for all estimates were widely overlapping and no significant difference between the estimates of the MID was observed for any subgroup.
Construct validity of the inter-patient method was supported by our finding that subjective comparison ratings and arithmetic differences correlated significantly ($r = 0.20$ to $0.22$ for the three outcome measures). The magnitude of correlation was less than in studies using the inter-patient method with other indices: 0.41 for the HAQ,$^{138}$ 0.29 for FEV-1,$^{140}$ 0.59 for the 6-minute walk,$^{141}$ and 0.20 to 0.44 for different dimensions of the CRQ. Perhaps the communication difficulties and other problems manifested by some of our participants hindered presentation of all relevant information during a short meeting. In some of the other inter-patient MID studies, participants were members of a rehabilitation programme and had the opportunity to observe one another on multiple occasions rather than at a single meeting.

As we anticipated, correlations for disease specific issues were higher than for general issues. The correlation coefficient for arithmetic differences versus subjective comparisons for head and neck specific issues was 0.33, as compared with 0.15 for the more general issue of social well-being. Differences in scores for all except the emotional subscale of FACT-H&N correlated significantly with their corresponding subjective comparison ratings. We hypothesize that emotional well-being may be a private issue which is difficult to share with others during a brief encounter.

We hypothesized that the MID, expressed as a percentage of the maximum instrument range, might be smaller for a multi-item index as opposed to a single item measure. Our data did not show such a relationship, with all of our outcome measures showing MID estimates in the 4 to 6% range for positive MID and the 6 to 14% range for negative MID. Previous estimates of MID have fallen within a similar range, suggesting that patients may be able to detect a constant proportional change on different measures. This contrasts with statistical methods of estimating MID, which may give smaller estimates for multi-item scales.
5. Discussion

Cohen\textsuperscript{146} has suggested that the MID is about 20\% of the standard deviation. Applying these criteria to our data gave statistical estimates of MID of 3 units for FACT-H\&N (2\% of the total instrument range), 0.03 units for TTO (3 \% of the instrument range), 0.02 units for DATE (2\% of the instrument range) and 3 units for the Karnofsky scale (3\% of the instrument range). The differences which were noticeable to our patients were larger than this statistical method would suggest.

As previously noted,\textsuperscript{140} estimates of MID do not apply uniformly to all individuals. To be detectable by all or nearly all patients, much larger changes need to be present. However, since estimates of the MID are useful mainly in the planning and interpretation of studies involving groups of patients, this observation in no way diminishes their importance. Individual diversity does not mitigate the scientific quest for group tendencies.

5.2 Subjective Comparison Ratings

A. Positive Skew

Subjective comparison ratings were skewed toward the positive end of the scale, indicating that most respondents considered their own well-being to be better than that of their partners. This finding of skewed comparisons has been reported previously in all studies using the inter-patient method.\textsuperscript{128, 138, 139, 140, 141} This tendency has been labeled optimism or ego bias.\textsuperscript{145} Interestingly, a corresponding tendency to rate oneself as improved has been seen using the intra-patient method.\textsuperscript{130}
5. Discussion

5.3 Utility Assessments

A. The “Perfect Health” Anchor State

Our concern that the perfect health anchor state might present difficulties was confirmed. Eleven patients claimed that they already had perfect health, despite having had treatment for laryngeal cancer. While clinically gratifying, this suggests some denial regarding the disease. Contrary to the premise of expected value decision making, an additional 18 patients did not want perfect health even if no penalty was incurred. Mean utility values for the entire group were significantly lower when these 29 patients were excluded from the analysis.

Previous studies have asked cancer patients to provide utilities for their personal QOL. Of 84 testicular and colorectal cancer patients, 13 expressed an unwillingness to even consider trading. Of these individuals, six considered the standard life expectancies that were offered to be unreasonable. Four stated reasons, including religious objection and adaptation, which were similar to those given by our patients for declining perfect health, two others found the task nonsensical, and one was unwilling to trade any time until her small children grew up.112 Hence, our patients do not seem more anomalous than do those in other studies, although real people may fall short of the economists’ ideal.

The anchor state may be less problematic when the TTO is used with hypothetical laryngeal cancer scenarios. No concerns were noted in 4 such studies.30, 67, 68, 69 However, Van der Donk70 reported that 6 of 39 respondents (10 larynx cancer patients, 10 floor of mouth cancer patients, 9 physicians and 10 healthy subjects) were unable to rate a T3 scenario with the TTO due to its complexity (2), because of religious convictions (1) or due to failure to complete the interview (3).
Preferences are sometimes difficult to express within the framework of expected value decision making. For example, "a patient may think that the decision in his or her case should be made according to universal considerations, such as those expressed in religion and philosophy, rather than according to the chaos of one's momentary impulses." The reasons stated by our patients for declining perfect health suggest that some patients were attempting to maximize expected value for society, rather than for themselves as individuals. Bursztajn's suggestion that utility assessments be customized to honour individual principles has not to our knowledge been attempted. The best approach may be to record patients' reasons for having difficulty and provide group means both with and without these patients.

B. Linearity of the Utility Function with Time

One potential concern in the use of the TTO is that it can be derived directly from expected value decision theory only under conditions of a linear utility function. Many studies have demonstrated a non-linear, convex utility function for most patients. This indicates that time in the near future is valued more highly than remote time. One implication for the use of the TTO is that patients with a longer remaining lifespan may be willing to trade a greater proportion of their time than those with a shorter remaining life. If this were the case, we would expect age to correlate with utility values, with younger patients indicating lower utilities than older ones.

On the other hand, younger patients might be expected to have better QOL due to less comorbid illness, and this should also be reflected by higher utility. We tested for these effects in two ways. Using regression, we found that age did not significantly predict TTO, DATE or FACT-H&N scores. We also did not observe a significant correlation of these instrument scores with age. These findings suggest that younger laryngeal cancer survivors do not have significantly better QOL than older patients, and also that in our sample, on average, the utility function was approximately linear.
5. Discussion

C. Advantages and Disadvantages of the DATE

We compared the lifetime TTO to a daily version (DATE), in which patients were asked to spend additional time each day in a dreamless, non-refreshing sleep in order to achieve perfect health. All patients seemed to understand the trade required, although one person could not complete either the TTO or the DATE. Another patient who could not complete the TTO (because he could not estimate his future lifespan and did not believe it was his place to make decisions about his time of death), was able to provide a utility using DATE. No patients indicated that the additional sleep, as described, would be desirable.

A greater number of patients were willing to trade at least some time on the DATE than on the TTO. This observation agrees with findings in the original study using DATE, and suggests that the daily version (DATE) may be easier for patients to relate to than the lifetime TTO. For at least one of our patients, the idea of trading time each day was acceptable, while dying sooner was not. It is possible that more patients were willing to trade with the DATE because “sleep” is less threatening than death.

Utility scores measured by the TTO and DATE methods were highly correlated (r = 0.59) in contrast to the moderate correlation observed in Buckingham’s report (r = 0.37). We found that DATE correlated significantly with FACT-H&N (r = 0.34) and Karnofsky score (r = 0.28), as did the TTO (r = 0.30, 0.34 respectively). Buckingham found lower correlations of the utility instruments with a modified version of SF-36 (DATE, r = 0.22; TTO, r= 0.12). In addition to the difference in the populations surveyed, one important distinction in our study is that utility assessments were all interviewer administered. Buckingham used a mail survey, and it is possible that the lower correlations in his study may have been due to respondents misunderstanding the task.
We used multiple regression to identify patient, tumour and treatment factors that predicted scores for the TTO and DATE. Number of relapses was a significant predictor of TTO scores, with better utility in patients who had never relapsed. For the DATE, disease of the glottic rather than supraglottic larynx predicted for higher utility. Both of these relationships are clinically plausible and in the expected direction. Our estimates of MID were similar for both measures, but as with the FACT-H&N and Karnofsky score, a more marked asymmetry between positive and negative MID was observed for DATE.

Our data supports the construct validity of the DATE, since it correlates highly with the TTO and to a lesser degree with QOL and performance status. However, it remains an experimental method. Future trials could test DATE versus TTO in other populations. A comparison of the DATE with the SG would also be important to explore its construct validity. Overall, we believe that the DATE offers a practical advantage over the TTO and should be incorporated into future trials in conjunction with other utility measures.

5.4 Health Questionnaires

A. Institutional Outcome and Health Questionnaire Scores

Overall, the QOL of our patients was high. The mean FACT-H&N score of 123/144 (86%) compares favourably with mean scores of 110/148 (74%),\(^{38}\) 104/148 (70%),\(^{62}\) 116/156 (74%)\(^{63,64}\) and 117/164 (71%)\(^{66}\) on earlier versions of the measure in studies of larynx cancer patients and mixed head and neck cancer patients. Clearly, within the head and neck cancer diagnosis, patients experience a significant range of QOL.
B. Test-retest Reliability of the FACT-H&N

For FACT-G, high test-retest correlation has been demonstrated \((r = 0.92)^{94}\) but this type of reliability has not yet been reported for the FACT-H&N. In our study, the FACT-H&N was administered to 98 patients on two occasions, with an interval ranging from one week to 3 months. Patients were clinically stable, and judgmental and statistical techniques suggested that little change took place between administrations. Our finding of a high correlation \((r = 0.82)\) suggests adequate reliability. The long interval between administrations of the instrument for some patients in our study likely accounts for much of the observed variation.

C. Gender Differences in QOL

We found gender to be the only significant predictor of scores for FACT-H&N, with men having a significantly higher mean score (125) than women (118). Interestingly, the magnitude of difference in mean scores for men versus women corresponds approximately with our estimates of MID. Significantly better QOL in males has been observed previously in population results for general QOL on both the EORTC-QLQ-C30\textsuperscript{78} and SF-36.\textsuperscript{148,149} The FACT-G has not been tested in a healthy population, and gender effects have not been reported in its validation studies.\textsuperscript{94,150} In mixed head and neck cancer patients, a trend toward lower FACT-H&N scores in women was observed \((p = 0.06)\) in a study of 47 patients,\textsuperscript{66} but no association was observed in a separate trial involving 50 patients.\textsuperscript{65} The small number of women in most laryngeal cancer trials makes gender differences in QOL difficult to measure.
5. Discussion

5.5 Relationship Between FACT-H&N and Health Utilities

A. Can Health Questionnaire Scores Predict Utilities

We observed a moderate but significant correlation between scores on the FACT-H&N and the two utility instruments (TTO, \( r = 0.30 \); DATE, \( r = 0.34 \)). A higher correlation coefficient was seen between the FACT-H&N and Karnofsky (\( r = 0.42 \)). This suggests that in our patients, health questionnaires and utility measures were not perceived identically. Functional considerations seem to have played a greater role in the questionnaire scores than in the utility scores.

The questionnaire we used was multi-dimensional but used an unweighted summary score. It is possible that the weighting of the functional dimension was greater than that perceived by our patients as indicated on a holistic measure such as the TTO or DATE. It is also possible that when applying a value to their well-being, our patients considered factors that are not included in the FACT-H&N, such as spiritual, financial, and philosophical concerns. That at least some of our patients were considering such issues is suggested by the reasons given by a few for declining perfect health. Additionally, not all of the patients who described their health as "perfect" when completing utility assessments indicated the maximum score on the FACT-H&N.

Previous studies that have compared health questionnaires and utility instruments have shown limited agreement between the two types of measures.\(^{151}\) In a study comparing the TTO to a disease-specific QOL instrument for myocardial infarction patients, correlation coefficients for the subscales ranged from 0.31 to 0.42.\(^{152}\) In asthma patients, SG utilities were weakly correlated with a disease-specific QOL instrument, the AQLQ (\( r = 0.19 \)) and with a generic QOL instrument, the Sickness Impact Profile (SIP) (\( r = 0.15 \)).\(^{153}\) Correlation of utility scores with
subscales scores for the generic RAND 36-Item Health Survey 1.0 (RAND-36) in patients with intermittent claudication\textsuperscript{154} was similarly modest, with coefficients ranging from 0.16 (pain) to 0.46 (mental health) for the TTO, and from 0.10 (pain) to 0.34 (social functioning) for the SG. In cancer patients, one study\textsuperscript{155} administered a computerized TTO and an unspecified health questionnaire to prostate patients, but did not report correlation between the two instruments.

Regression techniques have been used in an attempt to develop a predictive equation for utilities using generic health questionnaire results. Studies have shown that Medical Outcomes Study Short Form 36 (SF-36) scores explain between 25\textsuperscript{%}\textsuperscript{156, 157} and 43\textsuperscript{%}\textsuperscript{158} of the variance in TTO scores, similarly demonstrating that prediction of utility values from descriptive questionnaires remains an elusive goal.

5.6 Future Directions

Our finding of asymmetry between positive and negative estimates of MID should be confirmed in other patient samples. Replication of our study in patients with another type of cancer to demonstrate the reproducibility of our results could also assess the degree to which this observation is generalizable to all cancer patients.

While it appears to offer practical advantages, a direct comparison of the DATE to the criterion measure, the SG, should be undertaken. Comparison of the DATE to the TTO in patients with other cancer diagnoses would also help to establish its feasibility and validity. Other future studies could include a direct comparison of TTO instruments using different upper anchors, such as “perfect health” versus “your previous health.”
6. CONCLUSION

We conclude that the positive MID for these instruments in this setting is about 5% of the maximum instrument score, while the negative MID is significantly higher, at about 10% of the maximum instrument score. In our sample, a small improvement in well-being was relatively easy for patients to detect whereas a larger deterioration needed to be present before it was noticeable.

Comparison with previous studies suggests that across patient diagnoses, the magnitude of the MID, expressed as a percentage of maximum instrument score, may fall within a narrow range for a variety of QOL and functional status instruments. Confirmation of this range in a few more trials may permit estimation of the MID for new instruments without the necessity of carrying out an explicit MID study for each one.

Utilities and health questionnaires may be perceived differently by patients, suggesting that there is no simple way to convert one measure into the other. Direct measurement of utilities should be undertaken for decision analyses and the establishment of health policy.

Persistent challenges for utility assessment include the use of “perfect health” as an anchor, as well as the fact that not all individuals behave according to the theory of expected value decision making. The DATE instrument offers practical advantages over the TTO, but requires validation against the SG.

We found that overall QOL is high for survivors at our centre. The results of this study pertain mainly to laryngeal cancer patients treated with primary radiotherapy, and provide further support for our current treatment policy. We invite investigators in centres using a surgical approach to replicate our measurement of QOL, and we would welcome the opportunity to
6. Conclusion

compare results in the two cohorts. Utilities thus derived could be incorporated into a formal
decision analysis for patients with advanced laryngeal carcinoma.
7. TABLES

Table 1: Characteristics of Participants and Non-participants*

* non-participants includes all patients meeting study inclusion criteria who were not approached (24), were excluded (70), declined participation (124) or withdrew (23) from the trial. P-values calculated using the t-test for continual variables, ANOVA for nominal data and the Mann-Whitney U test for ordinal data.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Participants(n=98)</th>
<th>Non-participants(n=241)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>65</td>
<td>68</td>
<td>0.029</td>
</tr>
<tr>
<td>Female</td>
<td>14 (14%)</td>
<td>49 (14%)</td>
<td>0.955</td>
</tr>
<tr>
<td>Karnofsky - current</td>
<td>88</td>
<td>84</td>
<td>0.001</td>
</tr>
<tr>
<td>Karnofsky - initial</td>
<td>81</td>
<td>82</td>
<td>0.686</td>
</tr>
<tr>
<td>Larynx Subsite</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supraglottic</td>
<td>24 (24%)</td>
<td>49 (20%)</td>
<td></td>
</tr>
<tr>
<td>Glottic</td>
<td>74 (75%)</td>
<td>187 (78%)</td>
<td></td>
</tr>
<tr>
<td>Subglottic</td>
<td>0</td>
<td>4 (2%)</td>
<td></td>
</tr>
<tr>
<td>Transglottic</td>
<td>0</td>
<td>1 (0.4%)</td>
<td></td>
</tr>
<tr>
<td>Grade</td>
<td></td>
<td></td>
<td>0.450</td>
</tr>
<tr>
<td>1</td>
<td>21 (21%)</td>
<td>36 (15%)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>52 (53%)</td>
<td>144 (60%)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>10 (10%)</td>
<td>19 (8%)</td>
<td></td>
</tr>
<tr>
<td>unknown</td>
<td>15 (15%)</td>
<td>42 (17%)</td>
<td></td>
</tr>
<tr>
<td>Stage</td>
<td></td>
<td></td>
<td>0.322</td>
</tr>
<tr>
<td>I</td>
<td>54 (55%)</td>
<td>123 (51%)</td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>27 (28%)</td>
<td>67 (28%)</td>
<td></td>
</tr>
<tr>
<td>III</td>
<td>7 (7%)</td>
<td>27 (11%)</td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td>9 (9%)</td>
<td>24 (10%)</td>
<td></td>
</tr>
<tr>
<td>Primary radiation</td>
<td>96 (97%)</td>
<td>235 (97%)</td>
<td>0.335</td>
</tr>
<tr>
<td>Total laryngectomy</td>
<td>8 (8%)</td>
<td>23 (9%)</td>
<td>0.241</td>
</tr>
<tr>
<td>Number relapsing</td>
<td>9 (9%)</td>
<td>27 (11%)</td>
<td>0.724</td>
</tr>
<tr>
<td>Years since diagnosis</td>
<td>4.8</td>
<td>4.5</td>
<td>0.478</td>
</tr>
<tr>
<td>Years since 1st treatment</td>
<td>4.6</td>
<td>4.3</td>
<td>0.492</td>
</tr>
<tr>
<td>Years since last treatment</td>
<td>4.2</td>
<td>4.0</td>
<td>0.523</td>
</tr>
</tbody>
</table>
Table 2: Mean Quality of Life and Performance Status Scores

* N indicates number of patients for whom scores on each instrument were available. Observed minimum and maximum values are given, but the theoretical maximum was observed for each measure. For all instruments, the theoretical minimum value is 0.

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean score</th>
<th>S.D.</th>
<th>minimum</th>
<th>maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>FACT-H&amp;N</td>
<td>120</td>
<td>123.32</td>
<td>15.29</td>
<td>61</td>
<td>144</td>
</tr>
<tr>
<td>FACT-G</td>
<td>120</td>
<td>94.52</td>
<td>11.33</td>
<td>47</td>
<td>108</td>
</tr>
<tr>
<td>FACT-P</td>
<td>120</td>
<td>25.66</td>
<td>2.65</td>
<td>16</td>
<td>28</td>
</tr>
<tr>
<td>FACT-S</td>
<td>120</td>
<td>24.04</td>
<td>4.50</td>
<td>6</td>
<td>28</td>
</tr>
<tr>
<td>FACT-E</td>
<td>120</td>
<td>21.02</td>
<td>3.31</td>
<td>7</td>
<td>24</td>
</tr>
<tr>
<td>FACT-F</td>
<td>120</td>
<td>23.80</td>
<td>4.11</td>
<td>11</td>
<td>28</td>
</tr>
<tr>
<td>FACT-H</td>
<td>120</td>
<td>28.62</td>
<td>5.48</td>
<td>12</td>
<td>36</td>
</tr>
<tr>
<td>TTO</td>
<td>112</td>
<td>0.914</td>
<td>0.156</td>
<td>0.25</td>
<td>1</td>
</tr>
<tr>
<td>DATE</td>
<td>113</td>
<td>0.895</td>
<td>0.119</td>
<td>0.47</td>
<td>1</td>
</tr>
<tr>
<td>KARNOFSKY</td>
<td>120</td>
<td>88.47</td>
<td>8.42</td>
<td>70</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 3: Mean TTO and DATE Scores Excluding Patients Who Have Or Refuse “Perfect Health”

* N represents the number of patients for whom scores on each instrument were available, and who indicated neither that their current health was perfect nor that they would not want perfect health.

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>S.D.</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTO</td>
<td>84</td>
<td>0.878</td>
<td>0.174</td>
<td>0.25</td>
<td>1</td>
</tr>
<tr>
<td>DATE</td>
<td>85</td>
<td>0.853</td>
<td>0.166</td>
<td>0.47</td>
<td>1</td>
</tr>
</tbody>
</table>

7. Tables
Table 4: Reasons Stated for Declining Perfect Health (N = 18)

The concrete:  
"I don’t believe in miracles”  
"I’m a realist”  
"It would be unnatural”  
"I like to be natural”

The philosophical:  
"Perfect is not life”  
"I don’t want to be perfect every day”  
"I wouldn’t want my life to be perfect”  
"Perfect would be dull”  
"You have to have some challenges!”

The status quo:  
"Better what you know than what you don’t”  
"I don’t want to go monkeying with how I am”  
"I’m in a mental state where I just want to take what comes”  
"I’m quite happy as I am”

The altruistic:  
"I don’t want to be greedy”  
"I don’t want to ask for the moon. Others are worse off”

The religious:  
"I’m a religious person... it’s for God to decide”

The pragmatic:  
"If I had perfect health, I’d have to go back to work”

The student of life:  
"I’m the luckiest man on Earth. I have gained so much by surviving”

Table 5: Correlation Between QOL Measures*

* Spearman correlation coefficients. All p - values < 0.003.

<table>
<thead>
<tr>
<th></th>
<th>FACT-H&amp;N #1</th>
<th>FACT-H&amp;N #2</th>
<th>TTO</th>
<th>DATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>FACT-H&amp;N #1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FACT-H&amp;N #2</td>
<td>0.824</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TTO</td>
<td>0.297</td>
<td>0.394</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DATE</td>
<td>0.337</td>
<td>0.361</td>
<td>0.588</td>
<td></td>
</tr>
<tr>
<td>KARNOFSKY</td>
<td>0.422</td>
<td>0.498</td>
<td>0.337</td>
<td>0.283</td>
</tr>
</tbody>
</table>
Table 6: Correlation Between Arithmetic Differences on FACT-H&N Subscales and their Corresponding Subjective Comparison Ratings

* DIFF- variables refer to arithmetic differences on subscales of FACT-H&N (physical, social, emotional, functional, and head and neck subscales) as well as on the FACT-G and the entire FACT-H&N. COMP- variables refer to subjective comparison ratings of well-being (physical, social, emotional, functional, head and neck and overall well-being).

# R values indicate Spearman correlation coefficients, and p-values indicate the likelihood of each correlation arising by chance

<table>
<thead>
<tr>
<th>Comparison*</th>
<th>R#</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIFF-FACT-P vs COMP-P</td>
<td>-0.20</td>
<td>0.0001</td>
</tr>
<tr>
<td>DIFF-FACT-S vs COMP-S</td>
<td>-0.15</td>
<td>0.0001</td>
</tr>
<tr>
<td>DIFF-FACT-E vs COMP-E</td>
<td>-0.02</td>
<td>0.5000</td>
</tr>
<tr>
<td>DIFF-FACT-F vs COMP-F</td>
<td>-0.21</td>
<td>0.0001</td>
</tr>
<tr>
<td>DIFF-FACT-H vs COMP-H</td>
<td>-0.33</td>
<td>0.0001</td>
</tr>
<tr>
<td>DIFF-FACT-G vs COMP-overall</td>
<td>-0.19</td>
<td>0.0001</td>
</tr>
<tr>
<td>DIFF-FACT-H&amp;N vs COMP-overall</td>
<td>-0.19</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

Table 7: Correlation Between Differences in QOL Measures*

* Spearman correlation coefficients. All correlations are significant with p=0.0001

<table>
<thead>
<tr>
<th></th>
<th>DIF-FACT-H&amp;N</th>
<th>DIF-FACT-G</th>
<th>DIF-TTO</th>
<th>DIF-TTO</th>
<th>DIF-KARN</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIF-FACT-H&amp;N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIF-FACT-G</td>
<td>0.940</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIF-TTO</td>
<td>0.365</td>
<td>0.309</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIF-TTO</td>
<td></td>
<td></td>
<td>0.337</td>
<td>0.498</td>
<td></td>
</tr>
<tr>
<td>DIF-KARN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.333</td>
</tr>
<tr>
<td>Comp-overall</td>
<td>-0.195</td>
<td>-0.192</td>
<td>-0.216</td>
<td>-0.215</td>
<td>-0.196</td>
</tr>
</tbody>
</table>
7. Tables

Table 8: Estimates of Positive and Negative MID*

* positive MID indicates the smallest favourable difference on each instrument which was noticeable to the average patient, and negative MID indicates the smallest unfavourable difference on each instrument which was noticeable to the average patient.

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Positive MID (95% CI)</th>
<th>Negative MID (95% CI)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>FACT-H&amp;N</td>
<td>6.22 (1.42-11.04)</td>
<td>9.31 (4.94-13.68)</td>
<td>0.02</td>
</tr>
<tr>
<td>FACT-G</td>
<td>4.37 (1.00-7.74)</td>
<td>6.32 (3.95-8.69)</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>TTO</td>
<td>0.050 (-0.011-0.111)</td>
<td>0.054 (0.009-0.099)</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>DATE</td>
<td>0.055 (0.018-0.092)</td>
<td>0.097 (0.068-0.126)</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>KARNOFSKY</td>
<td>3.83 (1.16-6.50)</td>
<td>7.09 (5.15-9.03)</td>
<td>&lt; 0.0001</td>
</tr>
</tbody>
</table>

Table 9: Positive and Negative MID for Subscales of FACT-H&N*

* determined by comparing mean differences in arithmetic score for each subscale with its matched subjective comparison rating, e.g. positive MID for physical subscale is given by the difference between the mean DIFF-FACT-P for patients rating themselves as "a little bit better" than their partners, minus the mean DIFF-FACT-P for patients rating themselves as "about the same" as their partners.

<table>
<thead>
<tr>
<th>Subscale</th>
<th>Positive MID (95% CI)</th>
<th>Negative MID (95% CI)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical</td>
<td>1.4 (0.44-2.4)</td>
<td>2.2 (1.2-3.1)</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Social</td>
<td>1.9 (0.17-3.6)</td>
<td>2.9 (0.08-5.8)</td>
<td>0.004</td>
</tr>
<tr>
<td>Emotional</td>
<td>0.56 (-0.40-1.5)</td>
<td>0.70 (-1.1-2.5)</td>
<td>0.23</td>
</tr>
<tr>
<td>Functional</td>
<td>1.4 (0.24-2.5)</td>
<td>4.2 (2.6-5.8)</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Head &amp; Neck</td>
<td>3.3 (1.6-4.9)</td>
<td>3.5 (1.4-5.6)</td>
<td>&lt; 0.0001</td>
</tr>
</tbody>
</table>
Table 10: Positive and Negative MID As Percentage of Total Instrument Score*

* Absolute values of MID from Table 8 are indicated as a percentage of maximum possible score on each instrument, eg. for FACT-H&N, positive MID is 6.22 / 144 = 4.3%

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Positive MID (95% CI)</th>
<th>Negative MID (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FACT-H&amp;N</td>
<td>4.3% (1.0-7.7)</td>
<td>8.6% (3.4-14)</td>
</tr>
<tr>
<td>FACT-G</td>
<td>4.0% (0.9-7.2)</td>
<td>7.7% (5.0-11)</td>
</tr>
<tr>
<td>TTO</td>
<td>5.0% (0-11)</td>
<td>6.0% (0-13)</td>
</tr>
<tr>
<td>DATE</td>
<td>5.5% (1.8-9.2)</td>
<td>14% (0-5.4)</td>
</tr>
<tr>
<td>KARNOFSKY</td>
<td>3.8% (1.2-6.5)</td>
<td>10% (7.5-13)</td>
</tr>
</tbody>
</table>
8. FIGURES

Figure 1: Plot of FACT-H&N #1 versus #2 for visual inspection of change. Application of the statistical method confirms that the point in black above the line meets our definition of an outlier, with $|\text{FACT-H&N #1} - \text{FACT-H&N #2}| = 50$. The point below the line did not meet our criteria for an outlier, since $|\text{FACT-H&N #1} - \text{FACT-H&N #2}| = 26$. 
Figure 2: Distribution of Difference Scores for FACT-H&N

N = 938

Frequency of occurrence of arithmetic differences between individuals in absolute values of FACT-H&N score. The range of possible scores is 0 to 144. The differences are symmetrical about 0.
Figure 3: Frequency Distribution for Subjective Comparisons. Top panels show subjective comparison ratings for physical and social well-being, middle panels, emotional and functional well-being, and bottom panels, head and neck specific and overall well-being.

For subjective comparison ratings, a rating of 7 means "much worse", 6 is "somewhat worse", 5 "a little bit worse", 4 "about the same", 3 "a little bit better", 2 "somewhat better" and 1 indicates "much better".
Figure 4: Comparison of Arithmetic Difference in FACT-H&N versus Subjective Comparison of Overall Well-being

N = 861
r = 0.19
p = 0.0001
Figure 5: Comparison of Arithmetic Difference in Time Trade-off versus Subjective Comparison of Overall Well-being

N = 823
r = 0.22
p = 0.0001
Figure 6: Mean difference in FACT-H&N for each level of subjective comparison of overall well-being. A rating of 7 means “much worse”, 6 is “somewhat worse”, 5 “a little bit worse”, 4 “about the same”, 3 “a little bit better”, 2 “somewhat better” and 1 indicates “much better”. Bars show 95% confidence interval and x denotes the mean.
Figure 7: Mean difference in TTO for each level of subjective comparison of overall well-being. A rating of 7 means “much worse”, 6 is “somewhat worse”, 5 “a little bit worse”, 4 “about the same”, 3 “a little bit better”, 2 “somewhat better” and 1 indicates “much better”. Bars show 95% confidence interval and x denotes the mean.
Figure 8: Mean difference in DATE for each level of subjective comparison of overall well-being. A rating of 7 means "much worse", 6 is "somewhat worse", 5 "a little bit worse", 4 "about the same", 3 "a little bit better", 2 "somewhat better" and 1 indicates "much better". Bars show the 95% confidence interval and x denotes the mean.
Figure 9: Mean difference on the H&N subscale of FACT-H&N for each level of subjective H&N comparison. A rating of 7 means "much worse", 6 is "somewhat worse", 5 "a little bit worse", 4 "about the same", 3 "a little bit better", 2 "somewhat better" and 1 indicates "much better". Bars show the 95% confidence interval and x denotes the mean.
Figure 10: Estimates of positive MID for FACT-H&N from subgroup analyses. Results from analyzing all contrasts are shown as the final entry. Bars show the 95% confidence interval and x denotes the mean.
Figure 11: Estimates of negative MID for FACT-H&N for subgroup analyses. Results from analyzing all contrasts are shown as the final entry. Bars show the 95% confidence interval and x denotes the mean.
Figure 12: Cumulative probability of a patient rating himself at least "a little bit better" for a given arithmetic difference in TTO. For example, on average, a patient with a TTO of 1.0 would have a 60% probability of rating himself as at least "a little bit better" than a partner whose TTO score was 0.9 (difference in TTO = 0.1).
9. REFERENCES


3. Torrance GW, Thomas WH, Sackett DL. A utility maximization model for evaluation of health care programs. Health Services Research 1972; 7: 118-33


9. References


41. Leroux-Robert J. A statistical study of 620 laryngeal carcinomas of the glottic region personally operated upon more than 5 years ago. Laryngoscope 1975; 85: 1440-52


44. Kirchner J, Owen J. Five hundred cancers of the larynx and pyriform sinus. Laryngoscope 1977; 87: 1288-303


46. Som M. Cordal cancer with extension to vocal process. Laryngoscope 1975; 85: 1298-307

9. References


69. Llewellyn-Thomas HA, Sutherland HJ, Thiel EC. Do patients’ evaluations of a future health state change when they actually enter that state? Med Care 1993; 31: 1002-12


9. References


84. Deleyiannis FW, Weymuller EA, Colter MD. Quality of life of disease-free survivors of advanced (stage III or IV) oropharyngeal cancer. Head Neck 1997; 19: 466-73


108. Loomes G. Disparities between health state measures: an explanation and some
9. References

implications. York, UK: Department of Economics, University of York, 1988


124. Wright JG. The minimal important difference: who's to say what is important? *J Clin Epidemiol* 1996; 49: 1221-22


9. References

142. Ferketich S. Focus on psychometrics: aspects of item analysis. Research in Nursing & Health 1991; 14: 165-168


151. Revicki DA, Kaplan RM. Relationship between psychometric and utility-based approaches to the measurement of health-related quality of life. Qual Life Res 1993; 2: 477-87


10. APPENDICES
Larynx
(ICD-O C32.0, 1, 2, C10.1)

Rules for Classification

The classification applies only to carcinomas. There should be histological confirmation of the disease.

The following are the procedures for assessing T, N, and M categories:

- **T categories**
  - Physical examination, laryngoscopy, and imaging

- **N categories**
  - Physical examination and imaging

- **M categories**
  - Physical examination and imaging

Anatomical Sites and Subsites

1. Supraglottis (C32.1)
   - (i) Suprahypoid epiglottis [including tip, lingual (anterior) (C10.1), and laryngeal surfaces]  
     - Epilarynx
       - (including marginal zone)  
       - Supraglottis excluding epilarynx
   - (ii) Aryepiglottic fold, laryngeal aspect
   - (iii) Arytenoid
   - (iv) Infrahypoid epiglottis
   - (v) Ventricular bands (false cords)

2. Glottis (C32.0)
   - (i) Vocal cords
   - (ii) Anterior commissure
   - (iii) Posterior commissure

3. Subglottis (C32.2)
Regional Lymph Nodes

The regional lymph nodes are the cervical nodes.

TNM Clinical Classification

T – Primary Tumour

TX  Primary tumour cannot be assessed
T0  No evidence of primary tumour
Tis  Carcinoma in situ

Supraglottis

T1  Tumour limited to one subsite of supraglottis with normal vocal cord mobility
T2  Tumour invades mucosa of more than one adjacent subsite of supraglottis or glottis or region outside the supraglottis (e.g., mucosa of base of tongue, vallecula, medial wall of piriform sinus) without fixation of the larynx
T3  Tumour limited to larynx with vocal cord fixation and/or invades any of the following: postcricoid area, pre-epiglottic tissues, deep base of tongue
T4  Tumour invades through thyroid cartilage, and/or extends into soft tissues of the neck, thyroid, and/or oesophagus

Glottis

T1  Tumour limited to vocal cord(s) (may involve anterior or posterior commissure) with normal mobility
T1a Tumour limited to one vocal cord
T1b Tumour involves both vocal cords
T2  Tumour extends to supraglottis and/or subglottis, and/or with impaired vocal cord mobility
T3  Tumour limited to larynx with vocal cord fixation
T4 Tumour invades through thyroid cartilage and/or extends to other tissues beyond the larynx, e.g., trachea, soft tissues of neck, thyroid, pharynx

Subglottis
T1 Tumour limited to subglottis
T2 Tumour extends to vocal cord(s) with normal or impaired mobility
T3 Tumour limited to larynx with vocal cord fixation
T4 Tumour invades through cricoid or thyroid cartilage and/or extends into other tissues beyond the larynx, e.g., trachea, soft tissues of neck, thyroid, oesophagus

N – Regional Lymph Nodes
NX Regional lymph nodes cannot be assessed
N0 No regional lymph nodes
N1 Metastasis in a single ipsilateral lymph node, 3 cm or less in greatest dimension
N2 Metastasis in a single ipsilateral lymph node, more than 3 cm but not more than 6 cm in greatest dimension; or in multiple ipsilateral lymph nodes, none more than 6 cm in greatest dimension; or in bilateral or contralateral lymph nodes, none more than 6 cm in greatest dimension
N2a Metastasis in a single ipsilateral lymph node, more than 3 cm but not more than 6 cm in greatest dimension
N2b Metastasis in multiple ipsilateral lymph nodes, none more than 6 cm in greatest dimension
N2c Metastasis in bilateral or contralateral lymph nodes, none more than 6 cm in greatest dimension
N3 Metastasis in a lymph node more than 6 cm in greatest dimension
M – Distant Metastasis

MX  Distant metastasis cannot be assessed
M0  No distant metastasis
M1  Distant metastasis

pTNM Pathological Classification

The pT, pN, and pM categories correspond to the T, N, and M categories.

pN0  Histological examination of a selective neck dissection specimen will ordinarily include 6 or more lymph nodes. Histological examination of a radical or modified radical neck dissection specimen will ordinarily include 10 or more lymph nodes.

G Histopathological Grading

See definitions on page 18.

Stage Grouping

<table>
<thead>
<tr>
<th>Stage</th>
<th>pT</th>
<th>pN</th>
<th>pM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage 0</td>
<td>Tis</td>
<td>N0</td>
<td>M0</td>
</tr>
<tr>
<td>Stage I</td>
<td>T1</td>
<td>N0</td>
<td>M0</td>
</tr>
<tr>
<td>Stage II</td>
<td>T2</td>
<td>N0</td>
<td>M0</td>
</tr>
<tr>
<td>Stage III</td>
<td>T1</td>
<td>N1</td>
<td>M0</td>
</tr>
<tr>
<td></td>
<td>T2</td>
<td>N1</td>
<td>M0</td>
</tr>
<tr>
<td></td>
<td>T3</td>
<td>N0, N1</td>
<td>M0</td>
</tr>
<tr>
<td>Stage IVA</td>
<td>T4</td>
<td>N0</td>
<td>M0</td>
</tr>
<tr>
<td></td>
<td>T4</td>
<td>N1</td>
<td>M0</td>
</tr>
<tr>
<td></td>
<td>Any T</td>
<td>N2</td>
<td>M0</td>
</tr>
<tr>
<td>Stage IVB</td>
<td>Any T</td>
<td>N3</td>
<td>M0</td>
</tr>
<tr>
<td>Stage IVC</td>
<td>Any T</td>
<td>Any N</td>
<td>M1</td>
</tr>
</tbody>
</table>

102
## Summary

<table>
<thead>
<tr>
<th>Larynx</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Supraglottis</strong></td>
<td></td>
</tr>
<tr>
<td>T1</td>
<td>One subsite, normal mobility</td>
</tr>
<tr>
<td>T2</td>
<td>Involving mucosa of more than one adjacent subsite of supraglottis or glottis or adjacent region outside the supraglottis; without fixation</td>
</tr>
<tr>
<td>T3</td>
<td>Limited to larynx with vocal cord fixation or invades postcricoid area, pre-epiglottic tissues, base of tongue</td>
</tr>
<tr>
<td>T4</td>
<td>Extends beyond larynx</td>
</tr>
<tr>
<td><strong>Glottis</strong></td>
<td></td>
</tr>
<tr>
<td>T1</td>
<td>Limited to vocal cord(s), normal mobility</td>
</tr>
<tr>
<td>T2</td>
<td>Supraglottis, subglottis, impaired cord mobility</td>
</tr>
<tr>
<td>T3</td>
<td>Cord fixation</td>
</tr>
<tr>
<td>T4</td>
<td>Extends beyond larynx</td>
</tr>
<tr>
<td><strong>Subglottis</strong></td>
<td></td>
</tr>
<tr>
<td>T1</td>
<td>Limited to the subglottis</td>
</tr>
<tr>
<td>T2</td>
<td>Extends to vocal cord(s) with normal/ impaired mobility</td>
</tr>
<tr>
<td>T3</td>
<td>Cord fixation</td>
</tr>
<tr>
<td>T4</td>
<td>Extends beyond larynx</td>
</tr>
<tr>
<td><strong>All Sites</strong></td>
<td></td>
</tr>
<tr>
<td>N1</td>
<td>Ipsilateral single ≤3 cm</td>
</tr>
<tr>
<td>N2</td>
<td>Ipsilateral single &gt;3 to 6 cm</td>
</tr>
<tr>
<td></td>
<td>Ipsilateral multiple ≤6 cm</td>
</tr>
<tr>
<td></td>
<td>Bilateral, contralateral ≤6 cm</td>
</tr>
<tr>
<td>N3</td>
<td>&gt;6 cm</td>
</tr>
</tbody>
</table>
APPENDIX B: Introductory Letter to Prospective Patients

June 3, 1998

Dear Patient

You will soon be attending the clinic for follow-up of cancer of the larynx (voice box). I would like to make you aware of a study we are doing to measure the quality of life of patients. I will explain the study to you in more detail at your visit. However, if you would prefer not to discuss this study, you may contact me ahead of time at (416) 946-2123. No decision about whether or not to participate is required right now.

Quality of life refers to issues of health that affect a patient’s life as a whole. It is assessed by asking questions about physical, social, emotional, and functional well-being. In this study, I would ask you to fill out a questionnaire and give permission for me to review your chart. You would also be interviewed about some other aspects of your life and preferences.

Patients in the study will also be asked to meet in a group and decide how they are doing compared to other larynx cancer survivors. There will be a chance afterward to talk about thoughts or feelings about the meeting.

This study will not involve any medical tests or special treatments. All information will be kept confidential. You are free to choose whether or not to participate, and may withdraw at any time.

Your involvement in this study may help us to weigh the effects of treatment. Most importantly, you could help us to provide better information to future patients facing cancer of the larynx. We hope you will consider participating.

Sincerely

Dr. J. Ringash, B.Sc., M.D., F.R.C.P(C)
For the Head and Neck Site Group:

Dr. P. Gullane
Dr. B. O'Sullivan
Dr. B. Cummings
Dr. J. Irish

Dr. F. Liu
Dr. D. Payne
Dr. J. Waldron
Dr. P. Warde
APPENDIX C: Patient Consent Form

Updated: May 6, 1998

Informed Consent Document:
Quality of Life and Utility Assessment for Laryngeal Cancer Patients

Person to contact about this study: Dr. Jolie Ringash (416)946-2123.

I have been asked to participate in a study to measure the quality of life of patients who have been treated for cancer of the larynx (voice box).

Quality of life refers to issues of health that have affected my life as a whole. It is assessed by asking questions about my physical, social, emotional, and functional well-being. In this study, I am being asked to fill out a questionnaire and give permission for the above named doctor to review my chart. I will also be interviewed about some other aspects of my life and preferences.

I will then be asked to attend a meeting with a group of other larynx cancer survivors. I will talk to each of them individually, then decide how I am doing compared to others I meet. There will be a chance afterward to talk about any thoughts or feelings I have had about the meetings. Altogether, the study will require between four and five hours of my time.

This study will not involve any medical tests or special treatment. I will continue with any management and follow-up recommended by my doctor regardless of this study. I am free to choose whether or not to participate, and I may withdraw at any time without prejudice to my care. I understand that all information about me will be kept confidential.

I agree to participate in this study.

Signature of patient: ___________________________ Date: ______________

Witness: ___________________________ Date: ______________

I, the undersigned, have fully explained the study to the above volunteer.

Signature of Investigator: ___________________________ Date: ______________
APPENDIX D: FACT-H&N Version 4

Below is a list of statements that other people with your illness have said are important. By circling one (1) number per line, please indicate how true each statement has been for you during the past 7 days.

**PHYSICAL WELL-BEING**

<table>
<thead>
<tr>
<th>Item</th>
<th>Not at all</th>
<th>A little bit</th>
<th>Somewhat</th>
<th>Quite a bit</th>
<th>Very much</th>
</tr>
</thead>
<tbody>
<tr>
<td>I have a lack of energy</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I have nausea</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Because of my physical condition, I have trouble meeting the needs of my family</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I have pain</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I am bothered by side effects of treatment</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I feel ill</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I am forced to spend time in bed</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

**SOCIAL/FAMILY WELL-BEING**

<table>
<thead>
<tr>
<th>Item</th>
<th>Not at all</th>
<th>A little bit</th>
<th>Somewhat</th>
<th>Quite a bit</th>
<th>Very much</th>
</tr>
</thead>
<tbody>
<tr>
<td>I feel close to my friends</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I get emotional support from my family</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I get support from my friends</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>My family has accepted my illness</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I am satisfied with family communication about my illness</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I feel close to my partner (or the person who is my main support)</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

Regardless of your current level of sexual activity, please answer the following question. If you prefer not to answer it, please check this box □ and go to the next section.

<table>
<thead>
<tr>
<th>Item</th>
<th>Not at all</th>
<th>A little bit</th>
<th>Somewhat</th>
<th>Quite a bit</th>
<th>Very much</th>
</tr>
</thead>
<tbody>
<tr>
<td>I am satisfied with my sex life</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>
By circling one (1) number per line, please indicate how true each statement has been for you during the past 7 days.

**EMOTIONAL WELL-BEING**

<table>
<thead>
<tr>
<th>Statement</th>
<th>Not at all</th>
<th>A little bit</th>
<th>Somewhat</th>
<th>Quite a bit</th>
<th>Very much</th>
</tr>
</thead>
<tbody>
<tr>
<td>I feel sad</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I am satisfied with how I am coping with my illness</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I am losing hope in the fight against my illness</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I feel nervous</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I worry about dying</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I worry that my condition will get worse</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

**FUNCTIONAL WELL-BEING**

<table>
<thead>
<tr>
<th>Statement</th>
<th>Not at all</th>
<th>A little bit</th>
<th>Somewhat</th>
<th>Quite a bit</th>
<th>Very much</th>
</tr>
</thead>
<tbody>
<tr>
<td>I am able to work (include work at home)</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>My work (include work at home) is fulfilling</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I am able to enjoy life</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I have accepted my illness</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I am sleeping well</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I am enjoying the things I usually do for fun</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I am content with the quality of my life right now</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>
By circling one (1) number per line, please indicate how true each statement has been for you during the past 7 days.

### ADDITIONAL CONCERNS

<table>
<thead>
<tr>
<th>Statement</th>
<th>Not at all</th>
<th>A little bit</th>
<th>Somewhat</th>
<th>Quite a bit</th>
<th>Very much</th>
</tr>
</thead>
<tbody>
<tr>
<td>I am able to eat the foods that I like.</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>My mouth is dry.</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I have trouble breathing.</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>My voice has its usual quality and strength.</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I am able to eat as much food as I want.</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I am unhappy with how my face and neck look.</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I can swallow naturally and easily.</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I smoke cigarettes or other tobacco products.</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I drink alcohol (e.g. beer, wine, etc.).</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I am able to communicate with others.</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I can eat solid foods.</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>
APPENDIX E: Time Tradeoff (TTO) Semi-structured Interview Script

Updated: Oct. 1, 1998

Quality of Life and Utility Assessment for Laryngeal Cancer Patients

Script for semi-structured interview: T.T.O (Time Trade-Off) II

Hospital Number: ________________________________

Consider your present health. Think especially of any problems you may have as a result of your cancer, its treatment, or your other health conditions. Consider changes you have made in your life, including how you function and socialize. How you feel both physically and emotionally is important. Consider living with this health state indefinitely. During this time, your health would get neither better nor worse.

Now, considering your age and general health, about how long do you expect to live?

_________________________ x years ____________________________

Imagine I can give you a medicine which takes away all the problems related to your cancer and other health conditions. This medicine is easy to take and has no side effects. By choosing to take it, you would experience perfect health. This is a state of complete physical, mental and social well-being, not just the absence of disease or infirmity. Consider living with perfect health indefinitely. Imagine your health would remain the same every day.

I want you to make a choice between living for x years in your present health, the way you feel now, or x years in perfect health. In either state, your health would remain exactly the same up until the time you died. Which would you choose?

_________________________ ____________________________

Now, I want you to make a choice between living for x years in your present health, or 1 year in perfect health. Which would you choose?

_________________________ ____________________________

{the above question is repeated, changing the interval in perfect health by alternating decrements of 1 year and additions of 1 year, until the respondent is indifferent between x years in the present health state and y years in the perfect health state.}
APPENDIX F: Daily Active Time Exchange (DATE) Interview Script

Updated: Oct. 1, 1998

Quality of Life and Utility Assessment for Laryngeal Cancer Patients

Script for semi-structured interview: D.A.T.E (Daily Active Time Exchange) II

Hospital Number:_____________________________________________________

Please consider your present health. Think especially of any problems you may have as a result of your cancer, its treatment, or your other health conditions. Consider changes you have made in your life, including how you function and socialize. How you feel both physically and emotionally is important. Consider living with this health state every day.

Right now, how many hours on average would you be awake each day?

_________________________ x hours____________________________________

Imagine I can give you a medicine which takes away all the problems related to your cancer and other health conditions. This medicine is easy to take and has no side effects. By choosing to take it, you would experience perfect health. This is a state of complete physical, mental and social well-being, not just the absence of disease or infirmity. Consider living in perfect health every day.

I want you to make a choice between living in your present health, the way you feel now, or in perfect health. In either state, your health would remain exactly the same every day. In either state, you would be awake for x hours each day. Which would you choose?

_________________________________________________________________

Now, imagine that the medicine has one side effect. It will make you sleep more each day. This added sleep will be dreamless, and you will be no more refreshed than you are with your usual sleep. I want you to make a choice between two options. You may choose to live in your present health, being awake for x hours each day. Or, you may live in perfect health, but you will be awake for 1 hour each day. Which would you choose?

{the above question is repeated, changing the daily active time in previous health by decrements of 1 hour alternating with additions of 1 hour, until the respondent is indifferent between x active hours in the present health state and y active hours in the previous health state.}
APPENDIX G: Karnofsky Performance Status Scale

PERFORMANCE STATUS

<table>
<thead>
<tr>
<th>Definition</th>
<th>%</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Able to carry on normal activity and to work. No special care is needed.</td>
<td>100</td>
<td>Normal; no complaints; no evidence of disease</td>
</tr>
<tr>
<td></td>
<td>90</td>
<td>Able to carry on normal activity; minor signs or symptoms of disease</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>Normal activity with effort; some signs or symptoms of disease</td>
</tr>
<tr>
<td>Unable to work. Able to live at home, care for most personal needs. A</td>
<td>70</td>
<td>Cares for self. Unable to carry on normal activity or to do active work</td>
</tr>
<tr>
<td>varying amount of assistance is needed.</td>
<td>60</td>
<td>Requires occasional assistance, but is able to care for most of his needs</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>Requires considerable assistance and frequent medical care</td>
</tr>
<tr>
<td>Unable to care for self. Requires equivalent of institutional or hospital</td>
<td>40</td>
<td>Disabled; requires special care and assistance</td>
</tr>
<tr>
<td>care. Disease may be progressing rapidly.</td>
<td>30</td>
<td>Severely disabled; hospitalization is indicated although death not imminent</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>Very sick; hospitalization necessary; active supportive treatment necessary</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>Moribund; fatal processes progressing rapidly</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>Dead</td>
</tr>
</tbody>
</table>
APPENDIX H: Subjective Comparison Rating Form

Quality of Life and Utility Assessment for Laryngeal Cancer Patients

Conversation Record

A) The number on my nametag is #___________ (insert nametag number)

B) I am now talking with person #___________ (insert nametag number)

For each of the following statements, please circle the one (1) choice which best describes you compared to the person you just met.

C) Compared to this person, my physical well being during the past week was:

much better | somewhat better | a little bit better | about the same | a little bit worse | somewhat worse | much worse

D) Compared to this person, my ability to socialize during the past week was:

much better | somewhat better | a little bit better | about the same | a little bit worse | somewhat worse | much worse

E) Compared to this person, my emotional well being during the past week was:

much better | somewhat better | a little bit better | about the same | a little bit worse | somewhat worse | much worse

F) Compared to this person, my ability to function overall during the past week was:

much better | somewhat better | a little bit better | about the same | a little bit worse | somewhat worse | much worse

G) Compared to this person, my speech, eating and appearance during the past week were:

much better | somewhat better | a little bit better | about the same | a little bit worse | somewhat worse | much worse

H) Compared to this person, my overall quality of life during the past week was:

much better | somewhat better | a little bit better | about the same | a little bit worse | somewhat worse | much worse
APPENDIX I: Princess Margaret Hospital Clinical Trials Committee Approval

REGISTRATION OF A CLINICAL TRIAL AT OCI/PMH

TITLE: Quality of Life and Utility Assessment for Laryngeal Cancer Patients

APPLICANTS: J. Ringash, D. Redelmeier, B. O'Sullivan, A. Bezjak

STARTING DATE: March 1988 ESTIMATED COMPLETION DATE: accrual-Sept.'98

ESTIMATED NUMBER OF PATIENTS TO BE ENTERED FROM OCI/PMH: 100

1. Brief Description: Non-interventional cross-sectional study involving interviewing laryngeal cancer patients at the Princess Margaret Hospital to determine quality of life (using the FACT H&N questionnaire) and health-related utilities using traditional and daily time trade offs, and to correlate these measures with patient-perceived differences between individuals.

2. Required Services: Please identify any services that will be required for the clinical trial and an estimate of the annual costs that will be incurred that are additional to the costs associated with routine care.

The responsible Department Head should review the trial and initial this form to indicate that they are willing to accept the incremental costs. If no additional costs are anticipated for a department, please write "Not Applicable" under incremental costs.

<table>
<thead>
<tr>
<th>Service</th>
<th>Incremental Cost</th>
<th>Signature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnostic Imaging</td>
<td>Not applicable</td>
<td></td>
</tr>
<tr>
<td>Nuclear Medicine</td>
<td>&quot;&quot;</td>
<td></td>
</tr>
<tr>
<td>Oncologic Pathology</td>
<td>&quot;&quot;</td>
<td></td>
</tr>
<tr>
<td>Biochemistry</td>
<td>&quot;&quot;</td>
<td></td>
</tr>
<tr>
<td>Microbiology</td>
<td>&quot;&quot;</td>
<td></td>
</tr>
<tr>
<td>Pharmacy</td>
<td>&quot;&quot;</td>
<td></td>
</tr>
<tr>
<td>Biostatistics</td>
<td>&quot;&quot;*</td>
<td></td>
</tr>
</tbody>
</table>

* biostatistical support will be provided through I.C.E.S
3. Additional requirement for hospital admission or OPD visits:

Number of inpatient days 0 /yr

Chairman Inpatient Care Committee

Number of outpatient visits 0 /yr

Chairman Ambulatory Care Ctte.

Number of Day Care Visits 0 /yr

Chairman Ambulatory Care Ctte.

4. Budget: Please provide on a separate sheet a 1 page summary of the Protocol Trial budget indicating how the costs of the study are to be met.

-see appendix F

5. Space Requirements: Conference room in Head & Neck clinic during clinic time to interview patients (not currently in use for other purposes). Space in the early evening to hold 5 group meetings, each involving 20 patients and one facilitator. If evening facilities are not available at PMH, these meetings could be held at I.C.E.S.

6. Submission to the U. of T. Human Experimentation Committee:

Submission date: Feb. 9, 1998

Approval date: pending

7. Priority in division (to be completed by Department Head).

[ ] High [ ] Intermediate [ ] Low

[Signature]

Approved: [NAME]

(please print your name)
APPENDIX J: University of Toronto Oncology Human Review Committee Approval

University of Toronto

OFFICE OF RESEARCH SERVICES

PROTOCOL REFERENCE #3414

March 2, 1998

Dr. Brian O’Sullivan
Princess Margaret Hospital
610 University Ave.
Toronto, ON

Dear Dr. O’Sullivan:

Re: Research protocol entitled, “Quality of Life & Utility Assessment for Laryngeal Cancer Patients” (Dr. J. Ringash)

We are writing to advise you that the Oncology Human Review Committee composed of Drs. J. Ballinger, S. Gallinger, R. Myers, P. Warde, D. Warr, Professor B. Dickens and Ms. K. Tomson has granted approval to the above-named research study.

During the course of the research, any significant deviations from the approved protocol (that is, any deviation which would lead to an increase in risk or a decrease in benefit to human subjects) and/or any unanticipated developments within the research should be brought to the attention of the Office of Research Services.

Best wishes for the successful completion of your project.

Yours sincerely,

Susan Pilon
Executive Officer
Human Subjects Review Committee

cc: Dr. R. Ogilvie, Dr. P. Warde, Dr. J. Ringash, Princess Margaret Hospital